Enhancing Large Vision-Language Models with Ultra-Detailed Image Caption Generation

Yu Zeng[1] **Yukun Qi**[1] **Yiming Zhao**[1] **Xikun Bao**[1] **Lin Chen**[1]
Zehui Chen[1] **Shiting Huang**[1] **Jie Zhao**[2] **Feng Zhao**[1*]
[1]MoE Key Lab of BIPC, USTC [2]Huawei Noah's Ark Lab

Abstract

High-quality image captions are essential for improving modality alignment and visual understanding in Large Vision-Language Models (LVLMs). However, the scarcity of ultra-detailed image caption data limits further advancements. This paper presents a systematic pipeline for generating high-quality, ultra-detailed image captions, encompassing both pre-processing and post-processing stages. In the pre-processing stage, we classify and deduplicate images, extract visual information using expert tools, and leverage GPT-4o with structured prompts to generate initial captions. To enhance comprehensiveness, we introduce an expansion strategy based on Large Language Models (LLMs), defining eight descriptive dimensions to refine and extend captions, which serve as seed data for training a proprietary captioner model. In the post-processing stage, we incorporate human error-correction annotations and an active learning-inspired approach to refine low-quality samples. Using high-quality corrected data, we apply Direct Preference Optimization (DPO) and develop a critic-rewrite pipeline, training a sentence-level critic model to mitigate hallucinations. Experimental results demonstrate that our ultra-detailed captions significantly enhance LVLMs' perception and cognitive abilities across multiple vision-language benchmarks. The code and dataset are available at https://github.com/yuzeng0-0/UltraCaption.

1 Introduction

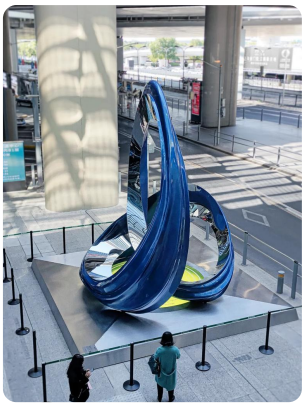
Large Vision-Language Models (LVLMs) (Bai et al., 2023; Chen et al., 2023a; Dai et al., 2023; Liu et al., 2023a; Luo et al., 2023; Ye et al., 2023; Chen et al., 2024a,b) have made significant progress in bridging the gap between language and vision, enabling tasks such as visual question answering and vision-language reasoning. However, the effectiveness of these models heavily depends on the quality of the image caption data used for pre-training. High-quality image captions are crucial for improving modality alignment, enhancing the understanding of visual content, and ensuring that models can generalize effectively across diverse visual and linguistic contexts. Unfortunately, the image caption datasets currently available for training LVLMs often lack the fine-grained details necessary to capture the complexity of images, particularly in terms of object attributes, object relationships, and intricate visual features.

The scarcity of ultra-detailed image caption data has become a key bottleneck in advancing LVLMs capabilities. Most existing caption datasets either lack sufficient visual detail or fail to provide comprehensive, detailed descriptions of the objects, attributes, and relationships within images (Chen et al., 2023b; Lai et al., 2023; Fan et al., 2023). As a result, LVLMs trained on such data may fail to align vision and language accurately, leading to suboptimal performance on tasks that require fine-grained visual reasoning. To address this challenge, this paper presents a comprehensive pipeline for generating high-quality, ultra-detailed image captions. The pipeline includes both pre-processing and post-processing stages, aimed at enhancing the quality of the generated captions and ensuring better alignment with the visual content.

In the pre-processing stage, we first classify and deduplicate the image data to build a multi-dimensional image dataset. We then leverage advanced visual expert tools (Zhang et al., 2023c; Liu et al., 2023c; PaddlePaddle) to extract rich visual information from the images and effectively integrate these visual information to generate high-quality image captions by prompting GPT-4o (Hurst et al., 2024). To further improve the completeness of the image caption data, we introduce an expansion strategy based on LLMs. Specifically, we design eight descriptive dimensions for the images, enabling the LLMs to expand and complement image details from multiple perspectives, ensuring a

* Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26715–26741
November 4-9, 2025



GPT4o: This image showcases a modern art sculpture located in a public space, possibly at the entrance of a transportation hub or a large building. The sculpture is primarily composed of blue and silver materials, featuring smooth shapes that embody a modern and abstract art style. The base of the sculpture is a square platform surrounded by railings to prevent people from getting too close. The design of the sculpture appears to be an intertwining of multiple curves and circular structures, with a smooth surface that reflects the surrounding environment. In the center, there is a bright yellow circular area that adds visual contrast and appeal. In the image, two people can be seen standing in front of the sculpture, with one of them taking a photograph. The background features some columns and architectural structures, indicating that this is an open and spacious area. In the distance, a few vehicles and pedestrians are visible, further confirming that this is a public place. The overall atmosphere is modern and filled with artistic flair.

Qwen2-VL 7B: This image shows a large blue metal sculpture located in a plaza. The sculpture consists of multiple curved ring structures, creating a sense of flow and dynamism. The surface of the sculpture reflects the surrounding environment, including buildings, trees, and the sky.
Beneath the sculpture, there is a yellow base that provides a striking contrast to the blue metal. Surrounding the sculpture are some black railings, likely intended to protect the sculpture and guide visitors. On the left side of the sculpture, some people can be seen taking photos or admiring the artwork. In the background, tall buildings and trees are visible, along with a road and traffic lights in the distance. The overall atmosphere appears very modern and artistic.

Ours: This image showcases a public art piece titled "Blue Time Machine" located within Shanghai Hongqiao Airport. The sculpture, designed by Tianjie Design Studio, is situated in an open area of the airport, surrounded by a modern and spacious environment. It consists of two interwoven blue metal rings, with a smooth surface that has a high gloss finish, reflecting the surrounding light and environment. The sculpture is placed on a square base, which is surrounded by black railings to ensure the safety of viewers. The railings appear to be made of metal, complementing the modern style of the sculpture. On either side of the sculpture, there are two thick columns in a light gray color, creating a striking contrast with the sculpture. One of the columns has a blue sign on its left side, with some white text written on it.
In front of the sculpture, two visitors are standing and admiring it. One is wearing a dark coat, while the other is in a light green coat, seemingly discussing or appreciating the artistic effect of the sculpture. In the background, other areas of the airport can be seen, including roads and distant buildings, indicating the airport's busyness and modernization.
The overall lighting is bright, with sunlight streaming through the architectural structures and casting unique shadows on the ground, enhancing the visual impact of the sculpture. The entire scene presents a public space filled with modernity and artistic ambiance.

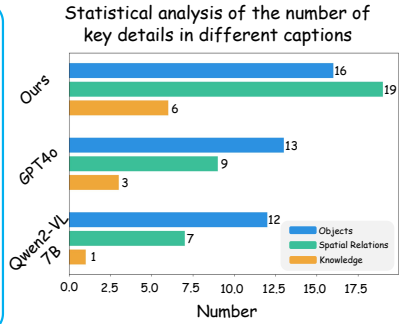


Figure 1: **We compare the performance of our method with the advanced GPT-4o and Qwen2VL-7B in image captioning.** For better visualization, objects are marked in blue, spatial relations in green, and knowledge are distinguished using yellow. The comparison clearly shows that our method captures finer details in the image more precisely and provides richer semantic understanding, further enhancing the quality of the image descriptions.

more holistic view. Building on the focus points expanded by the LLMs, we further prompt GPT-4o to extend the image captions to generate high-quality seed data, which are then used to train a proprietary image captioner model.

In the post-processing stage, we adopt an active-learning-like strategy to identify and correct bad samples generated by the model through human annotation. Using this high-quality, human-corrected data, we apply the DPO (Rafailov et al., 2024) alignment strategy to further enhance the image captioner model's performance. To make the most effective use of the human-corrected data, we also design a critic-rewrite pipeline. Specifically, we train a sentence-level critic model. For each caption, we decompose it into a series of atomic factual sentences and use the critic model to generate evaluative comments for each atomic sentence. Based on the original caption and the critic results for these atomic sentences, we rewrite the caption to further reduce hallucinations in the generated image descriptions. As shown in Figure 1, our method generates more accurate and comprehensive image captions across multiple dimensions compared to the advanced GPT-4o and the open-source model Qwen2VL-7B (Wang et al., 2024a).

Experimental results demonstrate that the image captions generated using our proposed pipeline significantly enhance the performance of existing LVLMs across various vision-language tasks and achieving better alignment between visual content and textual descriptions. In a nutshell, our contributions are as follows:

- We propose a powerful and scalable method for creating high-quality, ultra-detailed image captions, which is critical for enhancing the capabilities of LVLMs. Our method offers a promising solution to the limitations of current image caption datasets.
- We create a high-quality image caption dataset using the data generation pipeline proposed in this paper and validate its impact on enhancing the performance of LVLMs through data ablation experiments.
- We further validate the effectiveness of the proposed method in generating image captions through an image caption benchmark and manual quality analysis experiments, demonstrating enhanced caption performance and reduced hallucinations.

2 Related Work

2.1 LVLMs for Image-Text Data Enhancement

With the rapid development of Large Vision-Language Models (LVLMs) (Liu et al., 2023a; Luo et al., 2023; Ye et al., 2023; Zhang et al., 2023a,b; Zhu et al., 2023; Dai et al., 2023), research on image-text data enhancement has garnered increasing attention, focusing primarily on improving caption quality and vision-language alignment (Fan et al., 2023; Lai et al., 2023; Nguyen et al., 2023). LaCLIP (Fan et al., 2023) and VeCLIP (Lai et al., 2023) utilize LLMs to rewrite captions but are limited by issues of hallucinations. Models such as GPT-4V (OpenAI, 2023) can directly generate high-quality captions from images. Large-scale datasets like LAION (Schuhmann et al., 2021) and CC12M (Sharma et al., 2018), as well as synthetic caption generators like ShareGPT4V (Chen et al., 2023b), provide significant support for vision-language pretraining. DenseFusion (Li et al., 2024b) enhances caption quality by incorporating multimodal information during the data generation phase but remains largely confined to the data preprocessing stage. Despite progress, challenges persist in enhancing caption quality and reducing hallucinations. Future work should focus on further optimizing vision-language alignment.

2.2 Preference Optimization

Preference Optimization (PO) (Meng et al., 2024; Hong et al., 2024; Azar et al., 2024) is a key technique for advancing Large Language Models (LLMs) and Large Vision-Language Models (LVLMs). Methods like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) (Rafailov et al., 2024) use human preferences as reward signals to optimize model outputs, making them more aligned with human intent. In the multimodal domain, despite existing methods focusing on reducing hallucinations (Yu et al., 2023a, 2024), alignment optimization in complex image captioning scenarios remains challenging. To address this, Critic Models such as LLaVA-Critic (Xiong et al., 2025) and Prometheus-Vision (Lee et al., 2024) have emerged. These models can evaluate both visual and textual nuances, offering new ways to optimize alignment in complex multimodal tasks beyond single-task assessments.

3 Pre-processing Stage

3.1 Data Collection and Preparation

To create a high-quality dataset for precise vision-language perception, we have carefully assembled a diverse and multi-dimensional dataset with rich visual semantics to support the training of accurate and contextually aware image captioning models.

Data Sources. To maximize the diversity and comprehensiveness of the data, we have collected approximately 800K images from various sources, including the COCO dataset (Lin et al., 2014) for object detection, the SAM dataset (Kirillov et al., 2023) for image segmentation, and large-scale multimodal datasets commonly used in the field, such as Wukong (Gu et al., 2022), LAION (Schuhmann et al., 2021), CC3M (Sharma et al., 2018) and SBU (Ordonez et al., 2011). The data has been filtered and cleaned to remove corrupted, missing samples, and sensitive content.

Data Classification and Deduplication. To ensure data diversity and semantic richness, we design an image classification system with 7 primary categories and 22 secondary categories (as shown in appendix A) and finetune an image classifier to automate the categorization process. We also supplement underrepresented categories to balance the distribution. Finally, we apply deduplication based on image similarity, resulting in a dataset of 320K high-quality images.

3.2 Multimodal Information Fusion

Most LVLMs can generate image captions, but their quality is not guaranteed. Designed for general tasks, they often underperform in specialized areas like OCR and object detection compared to dedicated models. To address this, we develop a multimodal fusion pipeline that integrates visual information from specialized models with the advanced GPT-4o, to generate high-quality captions.

Extraction of auxiliary information. Due to the rapid advancements in the field of computer vision, many expert models in the visual domain can provide effective visual auxiliary information for image caption generation. We have carefully selected the following models to extract this visual auxiliary information:

• **Object Label Information.** We utilize RAM++ (Zhang et al., 2023c) to extract object label information from images and filter the original label vocabulary of RAM++ to remove labels that are not conducive to object detection, such as verbs

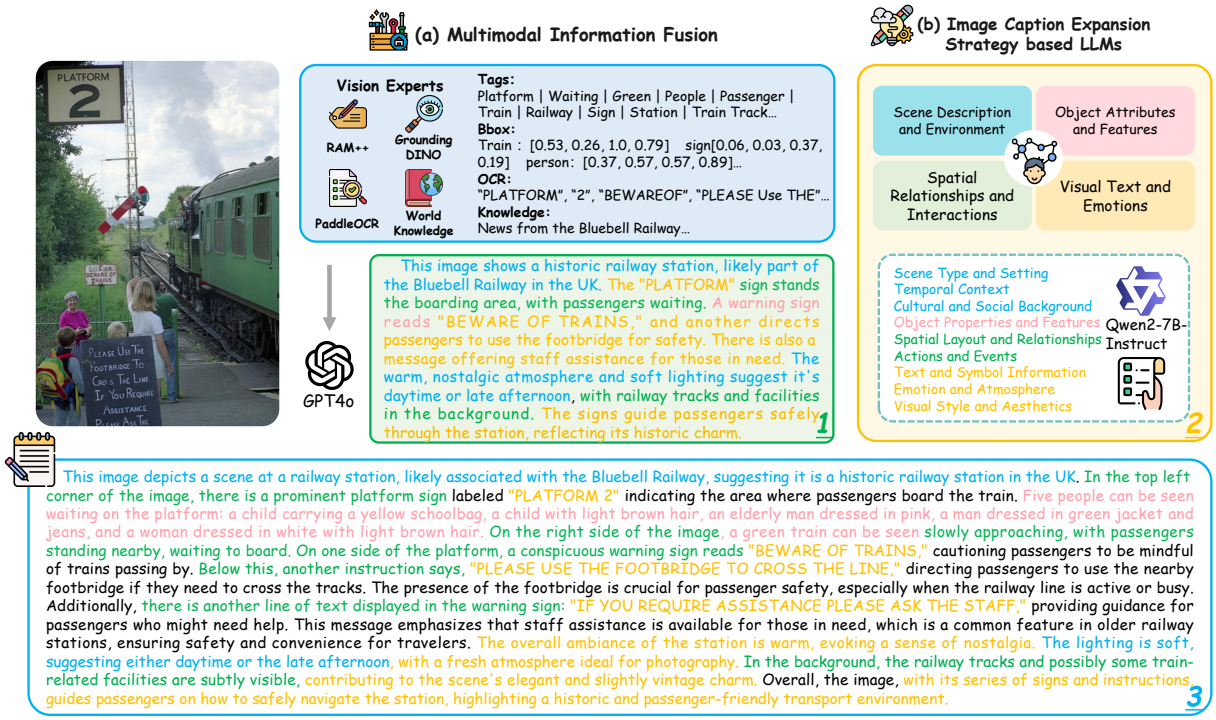


Figure 2: **Overview of pre-processing stage pipeline. (a)** We use advanced visual tools to extract detailed visual information and integrate it into GPT-4o via structured prompts to generate initial image descriptions. **(b)** We expand these descriptions using LLMs with eight defined dimensions (e.g., Scene Type, Object Properties, Spatial Layout, Text Information) to enrich details. This process generates 320k high-quality image caption seed data.

(e.g., "running"), adjectives (e.g., "bright"), and some background nouns (e.g., "sky", "ground").

• **Location Information.** We employ GroundingDINO (Liu et al., 2023c) to extract object detection box information from images. By leveraging the object labels extracted by RAM++, Grounding DINO recognizes the positions of the corresponding objects in the image and provides the respective bounding box coordinates for detection.

• **Textual Information.** We utilize PaddleOCR (PaddlePaddle) to extract textual information from images and filter out text with low confidence levels in the recognition results.

• **World Knowledge.** In most datasets, such as Wukong, LAION, and CC3M, images typically contain a raw descriptions related to the world knowledge of the image. Although these descriptions are very brief and lack fine-grained visual details, they contain rich world knowledge about the image.

3.3 Image Caption Expansion Strategy based LLMs

In Section 3.2, we prompt GPT-4o to generate relatively accurate image captions by integrating rich visual auxiliary information. However, considering that the generated captions may still overlook certain visual details, we introduce an image caption expansion strategy based on LLMs to further improve the completeness and comprehensiveness of the captions. Specifically, we first design eight descriptive dimensions for the images, enabling the LLMs to expand and complement image details from multiple perspectives, ensuring a more holistic view. Next, we input the image captions generated by GPT-4o, along with these eight dimensions, into the LLMs, allowing it to expand on the visual focus areas based on the details in the caption. Building on the focus areas identified by the LLMs, we further prompt GPT-4o to extend the image descriptions, effectively enhancing the completeness and comprehensiveness of the generated captions. In our implementation, we adopt Qwen2-7B-Instruct (Yang et al., 2024) as LLM. We provide concrete examples in Appendix C.

3.4 Captioner Model Training

As shown in Figure 2, we first construct preliminary image caption data by integrating visual auxiliary information, and then employ an image caption expansion strategy based on LLMs to effectively enhance the completeness and comprehensiveness of the generated captions. Through this approach, we build approximately 320K high-quality image

caption seed data. To break free from the reliance on costly proprietary models (such as GPT-4o) and achieve scalability in image caption data, we fine-tune a proprietary image captioner model using the Qwen2VL-7B (Wang et al., 2024a) model with the 320K high-quality seed dataset. The specific training details can be found in the appendix B.1. Through manual quality checks and image caption benchmark tests, our captioner model demonstrate performance comparable to GPT-4o, and even surpassed it in certain aspects.

4 Post-processing Stage

Previous studies, such as RLAIF-V (Yu et al., 2024), use automated preference data generation to enhance model performance by scoring or modifying responses with a divide-and-conquer strategy and multimodal models. However, we find this unreliable for image captioning tasks. Even powerful models like GPT-4o struggle with common issues such as counting in crowded scenes, object localization, and occluded scenes. Using GPT-4o to score or modify captions in these cases can introduce biases and destabilize the optimization process. Therefore, we emphasize the importance of incorporating human error-correction in the post-processing stage. In this section, as shown in Figure 3 we introduce human-corrected DPO alignment strategies and a critic-rewrite pipeline to improve the quality of image captions.

4.1 Preference Optimization

Human Error-correction Annotations. Based on the image captioner model trained in the pre-processing stage, we collect 200K diverse images using the image collection method outlined in Section 3.1 and generate image captions using the captioner model in pre-processing stage. To better identify low-quality image caption samples and make more effective use of human annotation resources, we adopt an active-learning-like strategy to filter out bad samples. Specifically, we rely on existing open-source critic models (e.g., LLaVA-Critic (Xiong et al., 2025)) to provide preliminary scores for the image captions. Although the scores provided by current critic models may not perfectly reflect the quality of the captions, they offer a rough estimation of quality that helps us filter out bad samples. Ultimately, we select 70K low-scoring samples for human error correction and annotation. Through manual error correction, we obtain 70K high-quality image caption preference pairs. Using these high-quality preference data, we apply DPO to further refine the image captioner model obtained in the pre-processing stage.

Improving Direct Preference Optimization (DPO). In our experiments, we observe that during DPO, as shown in Equation 1, the reward values for both chosen and rejected samples significantly decreased, leading to severe mode collapse in the model's output, characterized by the generation of large amounts of repetitive text. As the training data size increased, issues such as repetitive text became more frequent. We attribute this phenomenon to the discriminative nature of the DPO loss, which is essentially a classification loss. Relying solely on this loss may cause the model to iterate in the wrong optimization direction, thereby reducing its generative capability. Inspired by InternVL2-8B-MPO (Wang et al., 2024b), we introduce two auxiliary losses during the DPO process: normalized SFT loss and BCO loss, as shown in Equations 2 and 3, to maintain the model's generative ability and ensure the stability of preference learning.

$$\mathcal{L}_{DPO} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_c|x)}{\pi_0(y_c|x)} - \beta\log\frac{\pi_\theta(y_r|x)}{\pi_0(y_r|x)}\right), \tag{1}$$

$$\mathcal{L}_{SFT} = -\log\frac{\pi_\theta(y_c|x)}{|y_c|}, \tag{2}$$

$$\mathcal{L}_{BCO} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_c|x)}{\pi_0(y_c|x)} - \delta\right) - \log\sigma\left(-\left(\beta\log\frac{\pi_\theta(y_r|x)}{\pi_0(y_r|x)} - \delta\right)\right), \tag{3}$$

$$\mathcal{L} = \alpha_1\mathcal{L}_{DPO} + \alpha_2\mathcal{L}_{SFT} + \alpha_3\mathcal{L}_{BCO}, \tag{4}$$

where x represents the prompt, y_c represents the preference response after human error correction, y_r represents the original rejected response, β is the KL penalty coefficient, and the policy model π_θ is initialized from model π_0. δ represents the reward shift, calculated as the moving average of previous rewards to stabilize training. α_1, α_2 and α_3 represent the weights of each loss component.

As shown in Equation 4, by combining these two auxiliary losses (normalized SFT loss and BCO loss) with DPO, we effectively mitigate the issue of both chosen and rejected sample reward values decreasing simultaneously. Furthermore, the captioner model, after preference alignment, demonstrates improved performance. Additional training hyperparameters are provided in the appendix B.2. In appendix D, we conduct an in-depth analysis of the reward collapse issue for positive and negative

samples observed during the DPO process, based on our findings. Furthermore, we discuss potential solutions to this problem in detail.

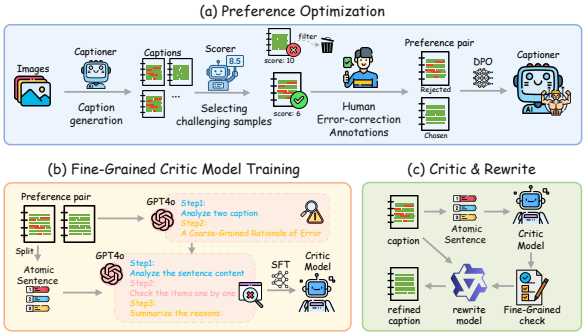


Figure 3: **Overview of post-processing stage pipeline. (a)** illustrates further preference optimization of the pre-processed image captioner model. **(b)** shows training of the fine-grained critic model. **(c)** depicts the critic-rewrite pipeline.

4.2 Fine-Grained Critic Model Training

In Section 4.1, we use a strategy similar to active learning to mine challenging samples and collect 70K high-quality human-corrected annotated data. Given that human-annotated data is both expensive and difficult to scale, we design a critic-rewrite pipeline to convert the human-corrected annotations into more granular sentence-level annotations, specifically for training a sentence-level critique model. By combining atomic sentence splitting and rewriting strategies, we further reduce hallucinations in the image captions. In the following, we will provide a detailed explanation of this process.

Coarse-grained Error Rationale Generation. We first input the human-corrected annotation data pair into GPT-4o and inform the GPT-4o which caption is correct and which is incorrect. This allows the GPT-4o to identify the differences between the two captions and generate a coarse-grained error rationale for the incorrect caption. The prompt template is provided in the figure 13.

Atomic Description Generation. To reduce the training difficulty of the critic model, we decompose the entire image caption into more granular atomic descriptions and perform critique evaluation at the atomic description level. Considering the contextual dependence in caption expressions, we ensure that each atomic description is as independent and specific as possible, while converting pronouns in the caption into explicit nouns to avoid ambiguity. Additionally, each atomic description should include as much relevant visual information as possible in a comprehensive manner. The prompt template is provided in the figure 12.

Fine-grained Critic Data Generation and Critic Model Training. We input both the image and a single atomic description into GPT-4o, along with the collected coarse-grained error rationale as a prompt, guiding GPT-4o to generate the evaluation process for that atomic description. To make the evaluation process more precise and controllable, we prompt GPT-4o to complete the task step by step. Upon receiving the atomic description, GPT-4o first needs to identify which details should be considered to evaluate all the visual information. Then, for each detail that needs attention, GPT-4o performs a comparison. Finally, GPT-4o summarizes the evaluation results. The prompt template is provided in the figure 14. In this process, we collect approximately 120K high-quality critic data and train a specialized sentence-level critic model based on Qwen2VL-7B. Training parameters and details are provided in the appendix B.3.

Rewrite of Captions. In the process described above, we obtain a fine-grained critic content for each atomic description. Subsequently, we input both the original image caption and the critic content for each atomic description into a LLM, prompting the model to rewrite the high-quality image caption. The prompt template is provided in the figure 15. We provide concrete examples in appendix E.

5 Experiments

Overview. We first introduce the key implementation details of the experiment and demonstrate that the high-quality ultra-detailed captions generated by our pipeline can effectively enhance the performance of LVLMs. Then, we conduct detailed ablation studies to validate the effectiveness of our pipeline in improving the quality of captions.

5.1 Implementation Details

Model and Data Setting. We use LLaVA-1.5 (Liu et al., 2023b) and LLaVA-NEXT (Liu et al., 2024) to validate the effectiveness of the high-quality, ultra-detailed captions generated by our proposed pipeline in enhancing the capabilities of Large Vision-Language Models (LVLMs). Specifically, we select CLIP-ViT-L/14-336 as the visual encoder, combined with Vicuna-v1.5-7B (Chiang et al., 2023) and LLaMA3-8B (AI@Meta, 2024) as the large language models. In the context of data, we meticulously select and collect approx-

Table 1: **Main Results.** Comparisons with state-of-the-art approaches on 9 vision-language evaluation benchmarks, including MME, MMB, MMB^{CN}, MMVet, SEED^{I}, POPE, SQA^{I}, GQA, and TextVQA. The results demonstrate that the high-quality image caption data generated by our method can bring significant and consistent improvements to Large Vision-Language Models (LVLMs). The best results are **bold** and the second-best are underlined.

Method	LLM	MME^{P}	MMB	MMB^{CN}	MMVet	SEED^{I}	POPE	SQA^{I}	GQA	TextVQA
Low-resolution Multimodal Large Language Models										
InstructBLIP	Vicuna-7B	-	36.0	23.7	26.2	53.4	78.9	60.5	49.2	50.1
QwenVL	Qwen-7B	-	38.2	7.4	-	-	56.3	67.1	59.3	**63.8**
QwenVL-Chat	Qwen-7B	1487	60.6	56.7	-	-	-	67.2	57.5	61.5
mPLUG-Owl2	LLaMA2-7B	1450	64.5	60.3	36.5	-	-	68.7	56.1	58.2
InternVL-Chat	Vicuna-7B	1525	-	-	-	-	86.4	-	62.9	57.0
LVIS-4V	Vicuna-7B	1473	67.1	-	34.6	-	84.0	68.4	62.6	-
ShareGPT4V	Vicuna-7B	1567	68.8	62.2	37.6	69.7	85.7	68.4	63.3	60.4
LLaVA-1.5	Vicuna-7B	1510	64.3	58.3	31.1	66.2	85.9	66.8	62.0	58.2
LLaVA-1.5(Ours)	Vicuna-7B	**1574**	70.7	62.8	37.9	70.1	87.3	70.1	**64.4**	61.7
LLaVA-1.5	LLaMA3-8B	1553	72.8	-	34.9	69.2	85.0	72.3	63.8	-
LLaVA-1.5(Ours)	LLaMA3-8B	1561	**73.8**	**68.9**	**39.6**	**73.0**	**87.6**	**74.1**	64.2	61.8
High-resolution Multimodal Large Language Models										
LLaVA-NEXT	Vicuna-7B	1519	67.4	60.6	43.9	70.2	86.5	70.1	64.2	64.9
LLaVA-NEXT(Ours)	Vicuna-7B	1528	68.8	60.8	**44.6**	72.0	**88.4**	71.4	65.0	69.9
LLaVA-NEXT	LLaMA3-8B	1591	72.6	69.0	42.1	72.7	86.8	73.4	64.8	65.0
LLaVA-NEXT(Ours)	LLaMA3-8B	**1596**	**74.4**	**69.8**	42.8	**75.0**	**88.4**	**78.7**	**65.6**	**71.0**

imately 1.5M images from multiple datasets, including COCO, LAION, CC3M, SBU, and SAM. Through the systematic data generation process proposed in this paper, we generate 1.5M high-quality image description data to validate the effectiveness of our dataset. Our training strategy is divided into the following three stages: **(1) Pre-alignment Stage:** In this stage, we use the 1.5M high-quality captions generated by our proposed pipeline as training data. The visual encoder and the LLM are frozen, and only the MLP is trainable. **(2) Pre-training Stage:** In this stage, we continue to use the 1.5M high-quality captions from the pre-alignment stage as training data. For LLaVA-1.5, following the approach of SharGPT4V (Chen et al., 2023b), we make the last 12 layers of the visual encoder, the MLP, and the LLM trainable. For LLaVA-NeXT, following the settings in (Li et al., 2024a), we make the entire model trainable. **(3) Instruction Finetuning Stage:** In this stage, we finetune the LLaVA-1.5 and LLaVA-NeXT models using the open-source LLaVA-mix-665K and LLaVA-NeXT-760K datasets, respectively. We make the MLP and the LLM trainable. The detailed training procedure is provided in the appendix B.4.

Evaluation Benchmarks. We evaluate the model's performance on 9 widely used visual understanding benchmarks, including MME (Fu et al., 2023), GQA (Hudson and Manning, 2019), TextVQA (Singh et al., 2019), SQA (Lu et al., 2022), MMBench (Liu et al., 2023d), MMBench-CN (Liu et al., 2023d), MM-Vet (Yu et al., 2023b), SEED (Li et al., 2023a), and POPE (Li et al., 2023b). These benchmarks cover a broad range of evaluation dimensions, such as visual reasoning, scene understanding, and scientific reasoning.

5.2 Main Results

The main result as Table 1. The experimental results indicate that the high-quality caption pre-training data generated by our pipeline significantly enhances the capabilities of LVLMs, demonstrating clear advantages on most visual-language benchmarks. Furthermore, compared to other image captioning methods, such as ShareGPT4V, our method provides more fine-grained and complex image descriptions by integrating additional visual auxiliary information. The introduction of a post-processing mechanism further reduces hallucinations in the captions, which contributes to improved alignment of the visual and linguistic modalities. Consequently, our method shows more pronounced advantages in fine-grained image understanding and hallucination benchmarks, such as TextVQA (Singh et al., 2019), MM-Vet (Yu et al., 2023b), and POPE (Li et al., 2023b).

5.3 Ablation Studies

To thoroughly validate the effectiveness of the pre-processing and post-processing pipeline, we design two different ablation experiments and used two distinct standards for validation.

Caption Benchmark. Traditional image caption evaluation metrics like CIDER (Vedantam et al.,

Table 2: **Results on the CompreCap Benchmark.** During the pre-processing stage, our approach significantly enhances the comprehensiveness of captions (measured by object and pixel coverage) compared to the baseline (**Note:** The baseline refers to Qwen2VL-7B, which we use to finetune the image captioner model during the pre-processing stage). During the post-processing stage, we can further eliminate hallucinations in the captions (measured by object and relation scores). Our method even surpasses the advanced GPT4o in some dimensions.

Method	Caption Length(words)	object coverage(%)	pixel coverage(%)	object score(0∼5)	relation score(0∼5)
Baseline (Qwen2-VL-7B)	143.66	71.97	57.31	2.56	2.71
Pre-processing Captioner	153.84	74.37	63.01	2.79	2.73
+ DPO	174.06	75.76	63.14	2.81	2.73
+ DPO + Critic-rewrite	171.65	75.96	**63.19**	**2.84**	2.77
GPT-4o	108.20	72.78	57.54	2.58	2.93
Human	133.61	**77.62**	59.58	2.78	**2.99**

Table 3: **Manual Quality Analysis.** Statistics of the equivalence rate between the output captions of different method and the reference captions, where 'Overall' represents the average equivalence rate across different dimensions. During the pre-processing stage, our method significantly enhances the comprehensiveness of captions compared to the baseline. Through the post-processing stage, we can further eliminate hallucinations in the captions.

Method	Completeness(%)	Hallucination(%)	Text Quality(%)	Overall(%)
Baseline (Qwen2-VL-7B)	27.0	52.5	98.3	59.3
Pre-processing Captioner	64.4	63.5	92.5	73.5
+ DPO	64.8	71.3	**99.5**	78.5
+ DPO + Critic-rewrite	**68.6**	**73.0**	97.8	**79.8**

2015), BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) rely on n-gram techniques, which are sensitive to caption styles and don't always reflect caption quality. Pre-trained CLIP models also struggle with longer captions. In contrast, CompreCap (Lu et al., 2024) is a structured benchmark that evaluates detailed image descriptions using a scene graph, focusing on object, attribute, and relationship matching. As shown in Table 2, we compare our method with advanced GPT-4o and human annotators on the CompreCap dataset and conduct an ablation study. The results demonstrate that our method not only matches but also surpasses GPT-4o in some dimensions, approaching the performance of human annotators. Moreover, the incorporation of pre-processing and post-processing significantly enhances the quality of generated captions and overall performance on CompreCap.

Although our method shows clear gains on CompreCap in terms of object coverage and caption completeness, its performance on the relation dimension lags slightly behind GPT-4o. Upon closer analysis, we found that many errors are caused by left–right perspective mismatches: the benchmark annotations follow the subject's viewpoint, while the models often default to the viewer's perspective. This misalignment leads to penalization even when the relational description is semantically consistent. We thus argue that the gap arises mainly from reference frame differences rather than a fundamental weakness in spatial reasoning, highlighting a well-known challenge in both image captioning and spatial VQA.

Manual Quality Analysis. To better validate the effectiveness of our caption generation pipeline, we introduce human evaluation to further assess the quality of the generated captions. Specifically, following the classification system proposed in Section 3.1, we select approximately 20 images from each category, forming an evaluation set of 411 images in total. First, we use GPT-4o to generate captions for all images in the evaluation set, which are then manually revised and supplemented to form reference captions. After the test model generates its captions, human quality inspectors compare each test caption with the reference caption and image, evaluating them across three main dimensions: **completeness** (whether the test caption omits any details from the image), **hallucination** (whether the test caption contains incorrect descriptions or fabricated content), and **text quality** (whether the test caption is grammatically correct, fluent, and free of major expression issues). The inspectors assess whether the test captions meet or exceed the quality of the reference captions in these dimensions. Finally, we calculate the equivalence rate, and the test results are shown in Table 3.

6 Conclusion

We propose an effective pipeline for generating high-quality, ultra-detailed image captions that significantly improve the performance of LVLMs. By integrating pre-processing and post-processing stage, we enhance the accuracy and comprehensiveness of captions, addressing the limitations of existing datasets. Our experiments show that these high-quality captions boost LVLMs performance across various vision-language tasks, achieving better alignment between visual content and text. This work provides a scalable solution for improving LVLMs capabilities and lays the foundation for future advancements in multimodal learning.

Limitations

Although the pipeline we proposed is capable of generating high-quality, ultra-detailed image caption data, there are still some limitations. Future optimization directions mainly focus on two aspects:

First, modality expansion. Currently, our pipeline primarily focuses on generating and optimizing captions for natural images. However, many methods and ideas have not yet been applied to other modalities, such as video data. Therefore, expanding the image caption generation pipeline to support more modalities will be a key area of future work.

Second, optimization and iteration of the image captioning model. Through the pre-processing and post-processing pipelines, we have made progress in improving the quality of image caption data. However, optimizing the performance of the image captioning model remains a priority. The current caption generation process is relatively complex, requiring the construction of a long pipeline for subsequent data expansion, which complicates practical deployment and application. In the future, we plan to integrate the different functions of the pipeline into a unified image captioning model, simplifying the entire process and making it more suitable for real-world deployment and application.

Acknowledgements

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

AI@Meta. 2024. Llama 3 model card.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *Preprint*, arXiv:2202.06767.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. 2023. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *Preprint*, arXiv:2401.06591.

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: What else influences visual instruction tuning beyond data?

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. 2024b. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *2407.08303*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Empirical Methods in Natural Language Processing*, pages 292–305.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision, Part V 13, pages 740–755.*

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. 2024. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. *arXiv preprint arXiv:2412.08614*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, pages 2507–2521.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*.

OpenAI. 2023. Gpt-4v(ision) system card.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *In Advances in Neural Information Processing Systems*, 24:1–9.

PaddlePaddle. Paddleocr. https://github.com/PaddlePaddle/PaddleOCR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:1–27.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2023c. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Image Classification Framework

To create a comprehensive and high-quality dataset that encompasses diverse image categories and rich visual semantics, we carefully curate and build a multi-dimensional dataset consisting of 7 main categories and 22 subcategories. During the pre-processing stage, We classify and deduplicate to select approximately 320K high-quality image samples, with the detailed category distribution shown in Figure 4.

Figure 4: The category distribution of high-quality image data during pre-processing stage.

B Training Setting Details

All training and inference experiments are conducted on 64 × Ascend 910b NPUs.

B.1 Captioner Model Training Details

We use the Qwen2VL-7B model and the pre-processed 320K high-quality seed dataset to fine-tune a proprietary image captioner model. The specific experimental settings are presented in Table 4.

B.2 Captioner Model DPO Training Details

We use 70K human-corrected high-quality image caption preference pairs to perform DPO training on the pre-processing stage captioner model, further enhancing its capabilities. The specific hyperparameter settings are shown in Table 5.

B.3 Sentence-level Critic Model Training Details

We use the Qwen2VL-7B model and 120K high-quality critic data to train a specialized sentence-level critic model. The specific experimental settings are presented in Table 6.

Table 4: Hyperparameter Details for Training the Captioner Model

Hyperparameter	Settings
DeepSpeed Stage	3
Warmup Ratio	0.03
Trainable Module	LLM
Epoch	1
LR Schedule	cosine
Learning Rate	1e-5
Image Resolution	1024
Batch Size	128
Cutoff Len	6144

Table 5: Hyperparameter Details for DPO Training of the Captioner Model

Hyperparameter	Settings
DeepSpeed Stage	3
Warmup Ratio	0.1
KL Penalty Coefficient β	0.1
Loss Weights α_1, α_2, α_3	0.8, 1.0, 0.2
Trainable Module	LLM
Epoch	1
LR Schedule	cosine
Learning Rate	5e-6
Image Resolution	1024
Batch Size	64
Cutoff Len	6144

B.4 LLaVA-1.5 and LLaVA-NEXT Training Details

The main training implementation details are described in the primary paper. In this section, we present a detailed explanation of the experimental setup used to train LLaVA-1.5 and LLaVA-NeXT for evaluating our dataset, as shown in Table 7 and Table 8.

C Implementation Details of the Image Caption Expansion Strategy

In this section, we present two case studies to demonstrate the implementation details of our

Table 6: Hyperparameter Details for Training the Critic Model

Hyperparameter	Settings
DeepSpeed Stage	3
Warmup Ratio	0.1
Trainable Module	LLM
Epoch	1
LR Schedule	cosine
Learning Rate	1e-5
Image Resolution	1024
Batch Size	64
Cutoff Len	6144

LLM-based image caption expansion strategy. As shown in Figure 5 and 6, we begin by prompting GPT-4o to generate an initial caption for the image, guided by multimodal information fusion. Then, using a structured prompt, we instruct Qwen2-7B-Instruct to expand upon this caption based on eight predefined image description dimensions.

Specifically, Qwen2-7B-Instruct takes the original caption and the dimension as input to generate a set of targeted expansion questions for each dimension. These generated questions, along with the original caption, are subsequently fed back into GPT-4o, prompting it to supplement potentially overlooked visual details and produce a more detailed and comprehensive caption.

D Issues in DPO Optimization and Potential Solutions Discussion

Why does the reward decrease for both positive and negative samples during DPO training? We first provide a theoretical explanation, then present a potential solution based on our experimental observations and the optimization objective of DPO.

$$\mathcal{L}_{DPO} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_c|x)}{\pi_0(y_c|x)} - \beta\log\frac{\pi_\theta(y_r|x)}{\pi_0(y_r|x)}\right), \tag{1}$$

The optimization objective of DPO, as shown in Equation 1. It is important to note that $\pi_0(y_c|x)$ and $\pi_0(y_r|x)$ remain a constant value during optimization. This leads to a potential issue in the DPO optimization process: specifically, when both $\pi_\theta(y_c \mid x)$ and $\pi_\theta(y_r \mid x)$ decrease simultaneously, but $\pi_\theta(y_r \mid x)$ decreases at a greater rate, the DPO loss objective can still be satisfied. However, the DPO loss lacks an additional constraint term to prevent this phenomenon.

In our experiments, we tracked the changes in $\pi_\theta(y_c \mid x)$ and $\pi_\theta(y_r \mid x)$ and confirmed that this issue does indeed occur, leading to a decrease in both positive and negative sample rewards. Additionally, at intermediate checkpoints, we observed a significant decline in the model's generation capability, characterized by excessive repetition in generated captions. These findings suggest that using DPO loss without additional constraints may lead the model to iterate in an unintended direction.

$$\begin{aligned}\mathcal{L}_{BCO} = &-\log\sigma\left(\beta\log\frac{\pi_\theta(y_c|x)}{\pi_0(y_c|x)} - \delta\right) \\ &-\log\sigma\left(-\left(\beta\log\frac{\pi_\theta(y_r|x)}{\pi_0(y_r|x)} - \delta\right)\right),\end{aligned} \tag{2}$$

How do the SFT and BCO losses help mitigate this issue?

1. Adding a supervised fine-tuning (SFT) loss term can prevent the degradation of the model's generation capability, ensuring that the model maintains strong generative performance throughout the optimization process.

2. As shown in Equation 2, the BCO loss introduces an anchor (δ) to encourage the policy model's probability of generating positive samples to exceed that of the reference model as much as possible. This mechanism helps mitigate the issue of declining positive sample generation probabilities and reward values.

In our experiments, we effectively prevent performance degradation caused by the model drifting in an erroneous direction by balancing the SFT loss, DPO loss, and BCO loss. This strategy also significantly reduces hallucinations in the model's outputs.

E Implementation Details of the Caption Critic-Rewrite Process

In this section, we illustrate the details of our Critic-Rewrite process through a representative case. As shown in Figure 7, given an image caption generated by Captioner model, we first decompose it into a series of atomic sentences, each of which independently refers to a specific object in the image and conveys a concrete description. We then evaluate each sentence using our fine-tuned critic model.

Table 7: Detailed Experimental Setup for LLaVA-1.5

Hyperparameter	Pre-aligning	Pre-training	Instruction Tuning
Data	1.5M caption(Ours)	1.5M caption(Ours)	LLaVA-mix-665K
Batch Size	256	256	128
Learning Rate	1e-3	2e-5(MLP, LLM), 2e-6(VE)	2e-5
LR Schedule		cosine decay	
LR Warmup Ratio		0.01	
Weight Decay		0	
Trainable Module	MLP	MLP, VE(last 12 layers), LLM	MLP, LLM
Epoch		1	
Optimizer		AdamW	
DeepSpeed stage		3	

Table 8: Detailed Experimental Setup for LLaVA-NEXT

Hyperparameter	Pre-aligning	Pre-training	Instruction Tuning
Data	1.5M caption(Ours)	1.5M caption(Ours)	LLaVA-NeXT-760K
Batch Size	256	256	128
Learning Rate	1e-3	2e-5(MLP, LLM), 2e-6(VE)	2e-5
LR Schedule		cosine decay	
LR Warmup Ratio		0.01	
Weight Decay		0	
Trainable Module	MLP	MLP, VE, LLM	MLP, LLM
Epoch		1	
Optimizer		AdamW	
DeepSpeed stage		3	

Finally, we rewrite the caption by integrating the original caption with the sentence-level evaluations from the critic model, resulting in a more accurate and refined image description.

F Prompt Template

In this section, we present all the prompts used in our paper, as illustrated in Figures 8 through 15.

G Visualizations of Our Dataset

In this section, to provide a more comprehensive demonstration of the high-quality image captions generated by our pipeline, we offer additional examples. Furthermore, to better showcase the performance of these high-quality captions across different ability dimensions, we employ different color annotations to highlight key elements. Specifically, objects are marked in blue, spatial relations in green, textual information in pink, and knowledge are distinguished using yellow, which facilitates a clearer presentation and understanding of the characteristics of each dimension.

H Computational Efficiency and Inference Cost Analysis

Our primary goal is to enable a smaller-scale open-source model (7B) to achieve captioning performance comparable to that of powerful proprietary models such as GPT-4o. We believe that the complexity of our method is a necessary trade-off to create a practical, efficient, and interpretable alternative for the open-source community.

In the process of extracting multimodal auxiliary information, we leverage traditional computer vision models such as GroundingDINO, RAM++, and PaddleOCR. These models are relatively lightweight, and their computational cost is negligible compared to the inference cost of large language models (LLMs). Within our caption generation pipeline, we first invoke a 7B-scale captioning model during the pre-processing stage to produce initial descriptions. These descriptions are then decomposed into atomic statements using the same 7B model. Each atomic statement is subsequently reviewed by a 7B critic model, which is

called approximately 7–8 times in total. Finally, a 7B model is used once more to rewrite the reviewed statements into the final caption. On average, the full pipeline involves around 10 invocations of a 7B model. However, due to the relatively small model size and the use of efficient inference frameworks (e.g., vLLM), the overall computational and inference overhead remains low.

This image captures an intense moment during an ongoing men's soccer match:
At the center of the frame, a player wearing a blue jersey is attempting to control or clear the ball with his leg. The ball is positioned right in front of his knee, and his movement appears powerful and swift. His body is slightly off the ground, with his left arm raised high and his right leg extended forward, seemingly about to strike the ball.
Right next to him, a player in a white jersey (wearing number 11) is making a determined effort to win the ball. He is leaning forward, eyes locked on the ball, with his right arm slightly extended—ready to make a move and challenge for possession.
On the left side of the image, another player in a blue jersey is watching the situation unfold. He appears poised to step in and engage at any moment.
On the right side of the frame, with his back to the camera, is another player in a white jersey. His shirt bears the text "WDMSC" and the number 15. He also seems to be moving toward the ongoing contest for the ball.
In the background, green grass, trees, and what appear to be benches or players' bags indicate that the match is taking place on an outdoor field.
The overall scene is full of energy, capturing a dynamic snapshot of competition and physical confrontation typical in soccer. The players are deeply focused and fully engaged, displaying wide, athletic movements that reflect a high level of intensity and sportsmanship.

Qwen2-7B-Instruct expansion questions for each dimension

Scene Type and Setting
Is this a professional match venue or an amateur field? Are there any distinctive elements such as spectator stands, boundary lines, a coach's area, or referees?

Spatial Layout and Relationships
Are the "left" and "right" sides in the image judged from the camera's perspective or from the players' point of view?

Object Properties and Characteristics
What is the design and color of the football?

Temporal Context
Can the season (spring, summer, autumn, or winter) or time of day (morning, afternoon, dusk) be inferred from the color of the grass and the angle of the lighting in the environment?

GPT-4o supplements potentially overlooked visual details based on the expansion questions generated by Qwen2-7B-Instruct for each dimension

This image captures an intense moment during an ongoing men's football match:
At the center of the frame, a player in a blue jersey is attempting to control or clear the ball with a mid-air motion. The ball is positioned just in front of his knee, and his movement is swift and forceful. His body is slightly off the ground, his left arm raised, and his right leg is powerfully extended forward, showcasing a highly competitive action.
Right next to him, a player in a white jersey—number 11—is aggressively pressing in. He leans forward, eyes fixed on the ball, with his right arm slightly extended, displaying a posture of imminent interception, full of urgency and focus.
On the left side of the image, another blue jersey player is intently observing the situation, ready to join the contest at any moment. On the right side, a white jersey player—number 15, with "WDMSC" printed on his back—is seen moving quickly toward the area of action, actively engaging in the game.
The references to "left" and "right" in this description are based on the camera's perspective (viewer's angle), not the players' orientation. The main action occurs slightly right of center, and the overall movement of the players is directed toward the right side of the frame. **The football itself has a white base with red spiral or curved patterns, differing from the traditional black-and-white panel design**.
In the background, you can see green grass, a few simple benches, and players' personal backpacks. The field is surrounded by unfenced woodland and tall grass, indicating that the match is **taking place on an outdoor amateur football field rather than a standardized professional stadium**.
Based on the natural setting, the grass appears dark green with some patches of yellowing, and the trees have relatively full but slightly sparse foliage—**suggesting early autumn**. The lighting in the photo is soft, and shadows are not sharply defined, indicating that **the photo was likely taken in the late afternoon or close to dusk, providing a warm and gentle natural light effect**.
The entire scene is full of dynamic energy, capturing a snapshot of high-speed confrontation, tension, and physical contact in the game. The players are in a state of intense focus and competitive engagement, demonstrating strong sporting spirit and passion—even in a non-professional setting.

Figure 5: Image Caption Expansion Strategy Based on LLMs: Example1.

This image depicts a bustling commercial pedestrian street, lined with numerous shops on both sides and filled with many pedestrians, creating a lively and vibrant atmosphere.
Timpson: The sign reads "The Quality Service People." It is a shop offering services such as shoe repairs, key cutting, and watch repairs. Promotional signs in the window advertise "WATCH REPAIRS" and "MICROCHIPPED CAR KEYS."
Robertson & Watt: This appears to be a jewelry or watch store, with products displayed in the window.
schuh: A shoe retail chain with a prominent green sign that stands out clearly.
There are also some partially visible stores that likely include clothing, beauty, or other retail businesses.
A clock is visible in the image, showing the time as approximately 3 o'clock.
In the center and right side of the image, many pedestrians can be seen—some walking, others browsing the shops. Most are dressed in relatively heavy clothing. The crowd includes elderly people, young adults, and children, reflecting a diverse urban population. Several people are carrying shopping bags, indicating that this is a commercially active street. The pedestrians' brisk pace conveys the fast rhythm of urban life.
The architecture is typical of older European style, with stone or brick buildings, tall windows, and shops on the ground floor. The upper floors may serve as offices or residential apartments.
The sky is clear and sunny, and the street surface reflects light, suggesting it may have recently rained or been cleaned.
In the lower-left corner of the image, the photographer's name "Swalo Photo" is marked.

Qwen2-7B-Instruct expansion questions for each dimension

Cultural and Social Context
Do the types of shops reflect the typical street culture of the UK or Western European countries?
Do Timpson and schuh suggest British domestic brands?

Spatial Layout and Relationships
Where is the clock located?

Temporal Context
Can the season (spring, summer, autumn, or winter) or the time of day (morning, afternoon, dusk) be determined?

Text and Symbol Information
Is the "Swalo Photo" credit accompanied by any other information such as copyright, date, location, or method of photography?

GPT-4o supplements potentially overlooked visual details based on the expansion questions generated by Qwen2-7B-Instruct for each dimension

This image shows a busy commercial pedestrian street, lined with numerous shops on both sides and bustling with pedestrians, creating a scene full of everyday life.
Timpson: The shop sign reads "The Quality Service People." It is a service store offering shoe repairs, key cutting, and watch repairs. Promotional posters in the window advertise "WATCH REPAIRS" and "MICROCHIPPED CAR KEYS."
Robertson & Watt: This appears to be a jewelry or watch store, with merchandise displayed in the window. Above its entrance, there is a large, vintage-style three-dimensional clock showing the time as approximately 3 o'clock.
schuh: A shoe retail chain with a bright green sign that is highly noticeable.
There are also some partially visible shops that likely include clothing, beauty, or other types of retail businesses.
In the center and right side of the image, many pedestrians can be seen—some walking, others browsing the shops. Most are dressed in relatively heavy clothing, suggesting **the photo was taken in a colder season. The lighting conditions indicate that it was likely taken in the morning or afternoon**.
The crowd includes elderly people, young adults, and children, reflecting a diverse urban population. Some people are carrying shopping bags, indicating a commercially active street. The pedestrians' brisk walking pace reflects the fast-paced rhythm of urban life.
The buildings exhibit typical old European architecture, mostly constructed of stone or brick, with tall windows. Shops occupy the ground floor, while the upper floors may be used as offices or residential apartments.
The sky is clear and sunny, and the street surface is reflective, possibly due to a recent rain or cleaning.
In the bottom left corner of the image, the photographer's name is marked: "Swalo Photo ©2011," indicating the photo was taken in 2011.
This photo was likely taken in a commercial district of a British city—possibly Glasgow, Edinburgh, or London. The street scene is highly characteristic and captures the everyday urban life of a typical British city.

Figure 6: Image Caption Expansion Strategy Based on LLMs: Example2.

This image captures a warm and delightful classroom scene, where several children are gathered around a wooden table, engaged in a learning or reading activity.
The most eye-catching part of the scene features three girls sitting side by side at the table, attentively flipping through books and materials, all sharing a light blue umbrella. In front of them are picture books, pencil cases, and other learning materials, suggesting they are likely involved in a picture book reading session or a classroom exercise. The girl on the far right is wearing a blue, white, and pink outfit with some English letters on the front, and her hair is styled in two ponytails. **The girl in the middle is dressed in yellow and is holding the umbrella with her right hand**, next to a blue object. The girl on the left is wearing a dress and **holding a lollipop in her right hand**.
On the left side of the image, a boy is leaning forward, holding a book in his hands, as if he is about to join the activity or is curiously observing the others.
The classroom is tidy and colorful. The wall is decorated with the words "Happy Birthday" and paper crafts such as stars and flowers, creating a cheerful atmosphere. **The bulletin board on the left reads "MENO"** and is covered with student work. **A wall clock shows the time as approximately 11:10 AM**. In the background, a cabinet neatly holds various stationery and art supplies. The overall environment is both creative and conducive to learning, reflecting a relaxed, playful, and enthusiastic classroom moment.

Decomposing the original image caption into sentence-level atomic descriptions

Description: Three girls are sitting at a wooden table, sharing a light blue umbrella.
Evaluation: Consistent with the image

Description: The girl on the far right is wearing a blue, white, and pink outfit with English letters on the front, and her hair is styled in two ponytails.
Evaluation: Consistent with the image

Description: The girl in the middle is wearing a yellow outfit, holding the light blue umbrella with her right hand, and there is a blue object next to her.
Evaluation: The girl in the middle is wearing an orange outfit, not yellow, and she is holding the umbrella with her left hand, not her right hand

Description: The girl on the far left is wearing a dress and holding a lollipop in her right hand.
Evaluation: The girl on the far left is holding a pen in her right hand, not a lollipop

Description: On the table in front of the three girls are picture books, pencil cases, and other learning materials.
Evaluation: Consistent with the image

Description: A boy appears on the left side of the image, leaning forward.
Evaluation: Consistent with the image

Description: The boy is holding a book in his hands.
Evaluation: Consistent with the image

Description: The classroom wall has the words "Happy Birthday" hanging on it.
Evaluation: Consistent with the image

Description: The wall is decorated with paper stars and paper flowers.
Evaluation: Consistent with the image

Description: The bulletin board on the left side of the classroom has the word "MENO" written on it.
Evaluation: The bulletin board on the left side of the classroom reads "MEMO"; the original sentence "MENO" does not match the image

Description: The clock on the classroom wall shows the time as approximately 11:10 AM.
Evaluation: The clock is not fully visible in the image, so the exact time cannot be determined.

Rewriting the caption by incorporating both the original image and the critic evaluations of each sentence

This is a warm and playful classroom scene. Several children are gathered around a wooden table, fully engaged in a learning or reading activity, creating a strong atmosphere of focus and curiosity.
At the center of the image, the most eye-catching feature is three girls who are attentively flipping through books and materials while sharing a light blue umbrella. The scene feels both whimsical and full of childhood innocence. In front of them are picture books, pencil cases, and other school supplies, suggesting they may be involved in a picture book reading session or a classroom exercise.
The girl on the far right is wearing a blue, white, and pink outfit with some English letters printed on it, and her hair is styled in twin ponytails. The girl in the middle is **wearing an orange outfit, holding the umbrella with her left hand**, with a blue object placed beside her. The girl on the far left is wearing a dress and **holding a pen in her right hand**, looking focused.
On the left side of the image, a boy is leaning slightly forward with a book in his hands, seemingly observing the girls' activity with curiosity or preparing to join in.
The classroom environment is clean and colorful. The wall is decorated with a "Happy Birthday" banner and handmade crafts such as paper stars and flowers, creating a joyful atmosphere. **The bulletin board on the left reads "MEMO"** and is filled with students' creative works. **Although a wall clock is visible in the image, it is partially obscured, making it difficult to determine the exact time**.
In the background, various stationery and art supplies are neatly arranged on shelves, contributing to a setting that is both creative and conducive to learning. Overall, the image captures a relaxed, childlike moment filled with enthusiasm for learning.

Figure 7: Implementation Details of the Caption Critic-Rewrite Process.

Prompt for Generating Captions Through Multimodal Information Fusion

You are an experienced image description expert, skilled in extracting details from images and transforming them into vivid and accurate written descriptions. Your task is to generate a detailed and coherent description based on the image and the provided visual assistive information, ensuring that all important elements of the image are covered. Please note that the visual assistive information may not be complete, and you will need to supplement the missing details based on visual clues from the image.

Requirements:
1. Object Appearance Description: Accurately describe the color, shape, quantity, size, function, and state of the objects in the image.
2. Behavior and Action: Capture the state, actions, and results of the behaviors of the objects in the image, describing both dynamic and static features.
3. Background and Environment: Describe the background environment, including the scene, lighting, weather, location, environmental details, dynamic/static elements, and overall atmosphere.
4. Text Information: Identify text in the image, including its content and location. You can use the provided OCR information to ensure it is accurately integrated into the description.

Visual Assistive Information:
Object Location Information: {Object Location Information}
Textual Information: {Textual Information}
Image World Knowledge: {Image World Knowledge}

Constraints:
1. Narrative Description: Please describe the image content in a coherent narrative format, avoiding a list structure.
2. Accuracy and Completeness: Ensure the description is accurate and as complete as possible, covering all important details.
3. Natural Flow: Keep the description natural and fluent, avoiding overly technical or mechanical language.

Goal:
To generate a detailed, accurate, and coherent description that covers all important elements in the image, including the appearance of objects, actions, background environment, and textual information. The description should be clear and easy to understand, allowing readers to accurately grasp the content of the image without actually seeing it.

Figure 8: Prompt for Generating Captions Through Multimodal Information Fusion.

Prompt for Expanding Image Details base LLMs

You're an excellent visual language assistant. You will receive an image description and some visual auxiliary information. To further enhance the comprehensiveness of the image description, you need to make reasonable speculations and further expansions based on the following dimensions. Please tell me what details I need to focus on to describe the image more comprehensively based on the original image description:

The multi-dimensional dimensions include:
1. Scene Type and Settings: including location, time, weather/light conditions, environment details, dynamic/static elements, and atmosphere.
2. Spatial Layout and Relationships: including relative position, height/distance/level/angle/direction, arrangement, interaction and connection of objects, etc.
3. Object Properties and Features: including color, shape, material, size, function, status, etc.
4. Text and Symbol Information: including text/symbol content, position, font style and color, language, meaning and function.
5. Emotion and Atmosphere: including emotional expression, atmosphere building, backstory, etc.
6. Temporal Context: including seasons, time periods, historical periods, weather conditions, day and night alternation, etc.
7. Cultural and Social Background: including identifying elements such as geographical location, historical background, religious beliefs, architectural styles, costumes and decorations, language and writing, art and music, social customs, food culture and festivals.
8. Visual Style and Aesthetics: including color collocation, light and shadow effect, overall design style, etc.

You need to dynamically select the dimensions you should pay attention to based on the specific image information, that is, select the appropriate dimensions based on the image content, including but not limited to the dimensions mentioned above. Note that the output needs to be presented in the form of questions.

The following is an example:
{example}

Auxiliary Information:
Object Location Information: {Object Location Information}
Textual Information: {Textual Information}
Image World Knowledge: {Image World Knowledge}

Original image description:
{Original Image Description}

To describe this image more comprehensively, you need to pay attention to the following details:

Figure 9: Prompt for Expanding Image Details base LLMs.

Example of Expanding Image Details Using LLMs

Object Location Information:
cat[0.15, 0.42, 0.28, 0.85] dog[0.55, 0.36, 0.78, 0.91] sofa[0.12, 0.60, 0.90, 0.95]
window[0.00, 0.00, 1.00, 0.40] boy[0.40, 0.50, 0.55, 0.85] girl[0.60, 0.50, 0.75, 0.85]
Textual Information:
None
Original Image Description:
This image shows a family living room with a cat and a dog in the room, with the cat lying on the sofa and the dog sitting on the carpet. There is a window in the room, from which you can see some sunlight coming in, bringing a warm light to the room. A boy and a girl sat on the ground, playing on the toys, looking happy. Some pillows are placed on the sofa, some simple cutlery on the table, the TV is off, and the background is simple and warm.

To describe this image more comprehensively, you need to pay attention to the following details:
1. Scene type and setting: Apart from sofas, dining tables and televisions, what other furniture or objects are key elements of this picture? What are the materials or colors of these items? Are there curtains installed outside the window?
2. Space layout and relationship: What is the layout between sofas, carpets, windows and other items? How do boys and girls relate to the location of sofas, televisions and windows? Are you close or far away from these items?
3. Object attributes and features: What is the color and material of the sofa? Is there anything particularly conspicuous on the sofa, such as a pillow or blanket? What color is the hair of cats and dogs? What is the color, shape and state of other items such as televisions, windows, dining tables, etc.?
4. Behaviors and Events: What are the movements of boys and girls? Are they playing attentively, talking, or exhibiting other behaviors (such as scrambling for toys)?
5. Cultural and social background: What may be the cultural background of the family from the perspective of the home environment? Does the style of furniture and objects reflect a particular social or cultural characteristic (e.g., modern minimalism or traditional decorative style)?

Figure 10: Example of Expanding Image Details Using LLMs.

Prompt for Further Expanding the Image Descriptions based on LLMs Result

You are an experienced image description expert, skilled in extracting details from images and transforming them into vivid, accurate text descriptions. Your task is to generate a detailed and coherent description of the image based on the image, visual auxiliary information, the original image description, and any additional details that require further attention. Ensure that all key elements of the image are included.

Visual auxiliary information:
Object Location Information: {Object Location Information}
Textual Information: {Textual Information}
Image World Knowledge: {Image World Knowledge}

Original image description:
{Original Image Description}

You need to pay further attention to the following details:
{LLM-based Expansion Result}

Ensure that the final image description contains all the details mentioned above and meets the following requirements:
1. Narrative description: Present the image content in a coherent, narrative format rather than as a list.
2. Accuracy and Completeness: Ensure that the description is accurate and as complete as possible, covering all of the aforementioned key details.
3. Naturally smooth: Maintain a natural and fluid description, avoiding overly technical or mechanical language.
4. Clear and easy to understand: Make the description clear and easy to understand, allowing readers to grasp the image's content without actually seeing it.

Goal:
Generate a detailed, accurate, and coherent description that covers all important elements of the image, including the appearance, behavior, background environment, and text information of the objects. The description should be clear and easy to understand, enabling the reader to fully comprehend the image's content through text alone.

Figure 11: Prompt for Further Expanding the Image Descriptions based on LLMs Result.

Prompt for Atomic Description Generation

You are an excellent visual language assistant. Please decompose the received image description according to the following requirements, generating a series of atomic descriptions.

Requirements:
1. Each atomic fact sentence should focus on the main visual details and specific facts, excluding non-visual information and subjective emotions.
2. The reference relationship in the atomic description must be clear, ensuring that each description can independently point to the object in the image, avoiding ambiguity that leads to unclear semantics.
3. Ensure that each atomic fact is as independent and specific as possible.

Image Description: {Image Description}

Figure 12: Prompt for Atomic Description Generation.

Prompt for Coarse-grained Error Rationale Generation

You are an excellent visual language assistant. You will receive an image and two text descriptions. One of the descriptions is the standard text description that accurately reflects the content of the image; the other description contains some inconsistencies with the image content. Please carefully compare these two text descriptions, focusing on the following aspects: the color, shape, quantity, size, function, and state of objects; the state, actions, and results of the objects' behavior; the details of the scene and environment; and the location and content of any text information. Based on the comparison, please identify the key reasons for the discrepancies between the second image description and the actual image content.

Human-corrected Image Description: {Human-corrected Image Description}
Original Image Description: {Original Image Description}

Figure 13: Prompt for Coarse-grained Error Rationale Generation.

Prompt for Fine-grained Critic Data Generation

You are an excellent visual language assistant. You will receive an image and a sentence. Please carefully check whether the sentence aligns with the image content, following the steps outlined below:

Task Steps:
1. Analyze the sentence content: List all the key points that need verification, such as: object quantity, color, positional relationships, status, interaction actions, textual information, etc. List all the visual details that need to be checked.
2. Check each point: Verify each key point description and compare it with the image content one by one. If discrepancies are found, explain in detail where the conflict lies, and clearly describe the actual content in the image.
3. Summarize the reasoning: Summarize the results of the check, indicating which parts of the sentence match or do not match the image content. Provide clear reasons for any inconsistencies, and specify the true visual content in the image.

Notes:
1. Detailed comparison: Check each point against the image content carefully, making sure no detail is overlooked.
2. Clear explanation: If the sentence matches the image content exactly, state this directly; if inconsistencies are found, describe the conflict clearly and specifically.
3. Complex content: If the image content is complex, break down the check into bullet points to ensure clarity and logical structure.
4. Proper language: Use clear, formal language to describe the issues found, providing adequate justification.

Additional Tips:
Common errors in the sentence may include: {Coarse-grained Error Cause}
Sentence to be checked: {Atomic Sentence}

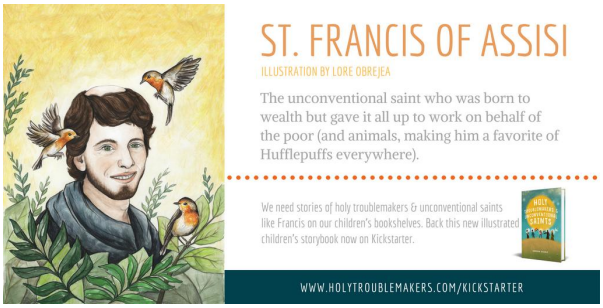
Figure 14: Prompt for Fine-grained Critic Data Generation.

Prompt for Rewrite of Captions

You are an excellent visual language assistant. You will receive an image and its original text description. The description contains some inconsistencies with the image content. I will provide you with a series of atomic sentences and their corresponding comments. Please rewrite the original image description based on these atomic sentences and comments, ensuring that the new description accurately reflects the image content.

Original Image Description: {Original Image Description}
Atomic Sentence and Critic: {Atomic Sentence and Critic}

Figure 15: Prompt for Rewrite of Captions.



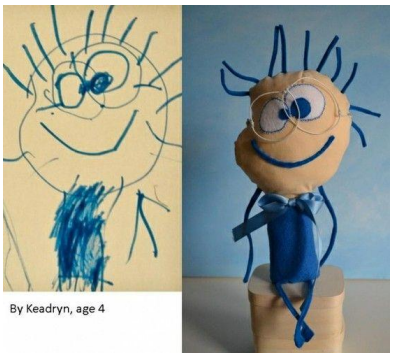
This image showcases a promotional poster for an illustrated storybook about St. Francis. On the left side, there is an illustration of a young man smiling, dressed in a robe, surrounded by green plants and several small birds. These birds are brightly colored, with orange and brown feathers that appear particularly vibrant in the sunlight.The background is a bright yellow, creating a warm feeling. On the right side is the text section, with the title "ST. FRANCIS OF ASSISI", illustrated by Lore Obreja. The text introduces the story of St. Francis: "This saint, born into a wealthy family, devoted his riches to the poor and formed bonds with animals, becoming a beloved figure in the Hecapuchin Academy worldwide". Below the text, there is encouragement for people to support a new illustrated children's storybook titled "Holy Troublemakers & Unconventional Saints", which is currently crowdfunding on Kickstarter. At the bottom of the poster, there is a URL link directing to the Kickstarter page. The overall design combines artistic illustrations and textual information, conveying a vibrant and hopeful atmosphere that draws readers' attention to the publication of this book.

In this image, we see the interior of a large white tent, with the ground covered in green artificial grass, creating an outdoor activity atmosphere. Inside the tent, there is a black chalkboard with the words "TATA TENT EVENT TIME" written in white chalk, along with the time "182 17:30" and the name "Yanis Varoufakis". The chalkboard is placed on a wooden easel, making it prominent and eye-catching.
In the background, multiple blue signs displaying "TATA TENT" are neatly arranged around the venue, forming an orderly area. The entire scene is well-lit, likely during the day, with bright lights inside the tent ensuring the smooth progress of the event. At the center of the image, there is a line of text stating "Copyright Athena Picture Agency Ltd," indicating that this is a copyrighted photograph. Overall, the image conveys a sense of preparation for an upcoming event, set in a spacious tent with a clean and organized environment.

In this image, we see a beautifully decorated blackboard wall located in an indoor environment. Various summer-themed patterns and texts are drawn on the blackboard with white chalk. In the center, it reads "KIABI", with the French phrase "la mode à petits prix" below, meaning "fashion at small prices". Next to it, "VACACIONES" is written in capital letters, meaning "vacation".
Above the blackboard, there is a drawing of a sun, symbolizing sunny weather. Beside the sun is a rainbow, adding a whimsical touch. On the right side, there is a drink cup with a straw, hinting at a summer beverage theme. In the lower left corner, there is a slice of watermelon next to a lifebuoy, further emphasizing the beach and vacation atmosphere. In the lower right corner of the blackboard, an open umbrella symbolizes shade and coolness. Nearby, there are several stars and a starfish, creating an oceanic feel. The lower left corner features another slice of watermelon, echoing the designs above. The background of the entire scene is a room with a colorful square carpet that adds vibrancy to the atmosphere. The walls are painted in simple colors, highlighting the patterns and texts on the blackboard. The overall ambiance is relaxed and cheerful, filled with the essence of summer.

Figure 16: The high-quality image caption data samples generated by our pipeline.

This image depicts a tranquil and beautiful countryside landscape, set against a vast valley backdrop where the sun is slowly rising in the distance, tinting the sky with hues of red and adding a warm golden glow to the entire scene. In the foreground, a lush green meadow is visible, with several sheep leisurely grazing, creating a harmonious and natural atmosphere.
On the right side of the painting, a small grove of trees can be seen, neatly arranged to form a natural barrier. In the distance, rolling hills are faintly visible on the horizon, adding depth and layers to the composition.
In the top left corner of the image, there is a text area stating, "10% of sales donated to Teesdale and Weardale Search & Mountain Rescue Team", indicating that a portion of the sales will be donated to the rescue team. The bottom right corner features the text "TEESDALE 2022 ANDY BECK images", marking this as the cover of the Teesdale calendar for 2022, photographed by Andy Beck. Overall, this image conveys a sense of tranquility and peace in the countryside through its soft colors and natural composition, while also expressing support and gratitude for the rescue team.



In this image, the left side features a child's drawing, while the right side presents a doll inspired by the artwork. The drawing was created by a four-year-old child, whose name is "Keadryn". The character in the drawing is a simple line art on a light yellow background, with large blue eyes and a smiling expression. The hair is outlined in thick blue lines, giving it a lively and cute appearance.
The doll on the right resembles the character from the drawing, using the same colors and style. The doll has a round body and a larger head, with similarly large blue eyes and a smiling mouth. Its hair is a deep blue, styled to appear as if a few strands are sticking up, adding a playful touch. The doll is dressed in blue clothing, with a blue bow tied around its waist, creating a lively and interesting overall design.
The background is a soft blue that complements the colors of the doll, creating a warm atmosphere. The entire composition is filled with childlike charm and creativity, showcasing a clever transformation from artwork to a tangible object.

In this photo, a hand is holding a ticket printed with "ALLEPPEY LIGHTHOUSE & MUSEUM TICKET", with a towering lighthouse in the background. The lighthouse features a striking red and white color scheme, topped with a circular observation deck. The nails on the hand are painted a deep red, and a silver ring is worn on one finger, which is naturally curved to display the ticket.
In the background, several palm trees can be seen, their leaves appearing especially lush and green in the sunlight, adding to the tropical ambiance. The sky is clear and pale blue, creating a bright and cheerful atmosphere. The ground is paved with a red brick pathway leading to the lighthouse, surrounded by green shrubs, contributing to a serene environment.
The ticket bears the words "150th ANNIVERSARY", indicating that this is a commemorative event for the lighthouse. Other details include the date and time "2019 15:31:43," as well as the ticket price "TOTAL 60.00". This photo captures a moment of a visitor while exploring the lighthouse, filled with the joy of travel and a sense of remembrance.

Figure 17: The high-quality image caption data samples generated by our pipeline.