

# Piculet: Specialized Model-Guided Hallucination Alleviation for MultiModal Large Language Models

Kohou Wang<sup>†</sup>[0009-0007-5863-2288], Xiang Liu<sup>†</sup>[0009-0003-2492-403X], Zhaoxiang Liu<sup>†,2</sup>[0000-0002-1267-0277], Kai Wang<sup>1,2</sup>[0000-0002-1171-0281], and Shiguo Lian<sup>†,2</sup>[0000-0003-4308-7049]

<sup>1</sup> AI Innovation Center, China Unicom, Beijing 100013, China

<sup>2</sup> Unicom Digital Technology, China Unicom, Beijing 100013, China

{wangzp103, liux750, liuzx178, wangk115, liansg}@chinaunicom.cn

**Abstract.** Multimodal Large Language Models (MLLMs) have made significant progress in bridging the gap between visual and language modalities. However, hallucinations in MLLMs, where the generated text does not align with image content, continue to be a major challenge. Existing methods for addressing hallucinations often rely on instruction-tuning, which requires retraining the model with specific data, which increases the cost of utilizing MLLMs further. In this paper, we introduce a novel training-free method, named Piculet, for enhancing the input representation of MLLMs. Piculet leverages multiple specialized models to extract descriptions of visual information from the input image and combine these descriptions with the original image and query as input to the MLLM. We evaluate our method both quantitatively and qualitatively, and the results demonstrate that Piculet greatly decreases hallucinations of MLLMs. Our method can be easily extended to different MLLMs while being universal.

**Keywords:** Multimodal Large Language Models · hallucinations · training-free.

## 1 Introduction

In recent years, there has been remarkable progress in the field of large-scale models, with the following being typical examples of this work: BERT [3], GPT-3[4], CLIP [5], DALL-E [6], etc. These works greatly promoted the development of Multimodal Large Language Models (MLLMs), an important branch of Artificial Intelligence. MLLMs' goal is to construct an artificial intelligence system capable of understanding and handling different modalities, such as image, text, audio, etc. The field of MLLMs has seen several landmark advancements, including

---

<sup>†</sup> These authors contributed equally to this work and should be considered co-first authors.

\* Corresponding authors

LLaVa, CogVLM, Minigt-5 et al. [7]–[9], they cast a profound impact on the improvement of MLLMs.

The rapid development of MLLMs has led to their widespread adoption for various applications, such as image captioning and visual question answering. Despite their impressive performance, MLLMs are still prone to generating hallucinations, where the generated text does not align with the image content. This issue significantly limits the practical applicability of MLLMs. As exemplified in Fig 1, the generated description of this image is not consistent with the truth:

1. there are 6 persons in the image, while the MLLM says **seven**;
2. there are 3 water bottles on the table, while the MLLM says **two**;
3. there is no **blackboard** on the left wall, only a **whiteboard**, while the model wrongly answers a **blackboard** after correctly answers a **whiteboard**.

Researchers have long been addressing the issue of hallucinations of MLLMs, and the mainstream methods can be divided into two kinds: training-based and training-free. Training-based methods usually collect or re-clean some datasets and retrain the models to decrease hallucinations of MLLMs. These kinds of methods, naturally, often require substantial manual intervention and are time-consuming. Moreover, given the substantial computational resources required for training large models, the economic cost of such methods is also quite considerable. As for training-free methods, current methods focus their emphasis on the postprocess of MLLMs. Given a user’s query, firstly the MLLMs will answer it as usual. Then the answer is thoroughly analyzed and corrected to decrease hallucination. These kinds of methods often utilize other rather large models apart from the MLLMs to be corrected, which is time-consuming and uneconomical.

Inspired by the phenomenon that humans use specialized tools to enhance their abilities, we propose a training-free framework, named Piculet, to enhance the input representation of MLLMs by leveraging multiple, specialized, small-scale, deep learning models to extract a description of visual information from the input image, i.e., we use multiple specialized models to guide MLLMs to generate more accurate results. Specifically, our Piculet utilizes the outputs from these small-scale deep learning models as external knowledge to enhance the MLLMs, thereby minimizing MLLMs’ propensity for hallucinations.

By combining the extracted description of visual information with the original image and query as input to the MLLM, we aim to improve the accuracy of the model’s output. Our method requires no retraining of MLLMs and no other rather large models, which is much faster and economical than all current available training-based methods and training-free methods. We evaluate the effectiveness of our method through comprehensive quantitative and qualitative experiments on the POPE [13], MME [14], and LLaVA-QA90 [15] datasets. The results and associated analyses indicate the superiority of this new paradigm. For instance, on the LLaVA-QA90 benchmark, our method largely boosts the accuracy of the baseline Qwen-VL-Chat [34] from 6.1 to 7.3 on a scale of 10.

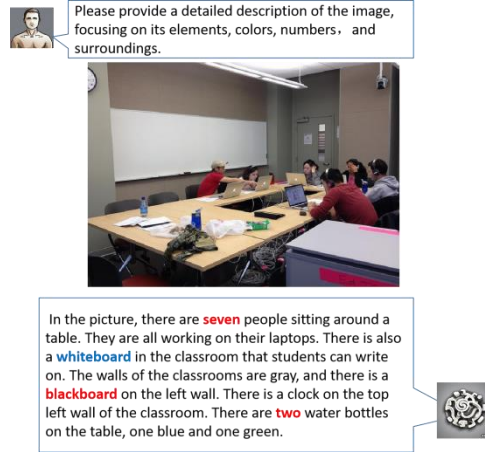


Fig. 1: Illustration of Hallucination of MLLMs. This MLLM generates descriptions of an image with wrong information, including the number of people and cups, and also, there is only whiteboard, not a blackboard on the left wall.

In summary, the main contributions are as follows:

- We proposed a training-free, pre-process framework named Piculet to reduce the hallucinations of MLLMs. To the best of our knowledge, we are the first to utilize a pre-process framework to tackle the visual hallucination problem.
- Our framework only requires one inference of the target MLLM and several other small deep learning models, which is economical and time-saving, and is plug-and-play in various different MLLMs. These small models' information is utilized as external knowledge to calibrate the MLLM.
- We evaluate our method on numerous datasets with other methods, and the results demonstrate the effectiveness and improvement of our method.

## 2 Related Work

### 2.1 MLLMs' Hallucinations

Despite the mushrooming of MLLMs, the problem of hallucination still hangs like the sword of Damocles: MLLMs occasionally generate content that diverges from the user input, contradicts previously generated context, or misaligns with established world knowledge. While the relatively usual normal deep learning models [18–21] output results of quite reliable credibility, hallucination puts the MLLMs at a disadvantage, users tend to use MLLMs more for fun rather than for professional needs, which is certainly not a good thing for MLLMs developed

for professional purposes. To address this challenge, existing mainstream works have primarily focused on two aspects: training-based and training-free.

## 2.2 Training-based Methods

For training-based methods, Gunjal et al. [10] introduced MHalDetect, a multi-modal hallucination detection dataset that can be used to train and benchmark models for hallucination detection and prevention. Liu et al. [11] addressed this issue by introducing the first large and diverse visual instruction tuning dataset, named Large-scale Robust Visual(LRV)-Instruction. Lu et al. [23] developed an evaluation module that automatically creates fine-grained and diverse visual question answering examples to assess the extent of agnosia in MLLMs comprehensively. They also developed a mitigation module to reduce agnosia in MLLMs through multimodal instruction tuning on fine-grained conversations.

These training-based methods, also well known as instruction-tuning, usually introduce a new dataset for retraining MLLMs, which requires significant computational resources and specialized data. These methods are also fairly time-consuming, considering that the inference of MLLMs is often a much longer time than traditional deep learning models.

## 2.3 Training-free Methods

As for training-free methods, Yin et al. [12] represents a typical method that requires no training of MLLMs while can directly correct the hallucinations. They emphasized main attention on the post-process stage of MLLMs, firstly they get an answer of a MLLM, then utilized auxiliary models' outputs to correct both object-level and attribute-level hallucinations, which was the first to apply a corrective manner to tackle the visual hallucination problem. Although their method, named Woodpecker, can reduce hallucinations by correcting the MLLM's answers, their method is a post-process framework, and still actually comprises three pre-trained rather large-scale models apart from the MLLM to be corrected, which are GPT-3.5-turbo [30], Grounding DINO [31] and BLIP-2-FlanT5 XXL [32]. Furthermore, the GPT-3.5-turbo is used 3 times in their processing pipeline. These models are not only time-consuming, but some are also proprietary, making them uneconomical with slow inference processes.

Compared to their approach, our method addresses the hallucination issue of MLLMs at its root and utilizes no other large language models apart from the MLLM to be corrected. Different from the Woodpecker method focusing on the post-process stage, our method focuses on the pre-process stage of MLLMs. Our method utilizes specialized, traditional small deep learning models to generate results describing factual information. These results, reorganized into a specific format, serve as supplementary descriptions and are input alongside the user's query and image into the MLLM, thereby enabling the model to generate correct answers directly by referencing additional factual information. Our specialized models only need to be run once during the processing pipeline, and the outputs of specialized models serve as external knowledge to calibrate the MLLM.

### 3 Method

Our method aims to address the hallucinations of MLLMs at its original source. We firstly utilize specialized traditional light-weight deep learning models to detect factual information of input image, then formulate these descriptions, which, alongside the user’s query and image, are input into MLLMs. MLLMs, given the formulated input, then generate results with reduced hallucinations. Our method utilizes these specialized models to generate factual external knowledge apart from the single input image, which provides a reliable basis for decision-making in the outputs of the MLLMs. We will introduce these steps in detail in sequence.

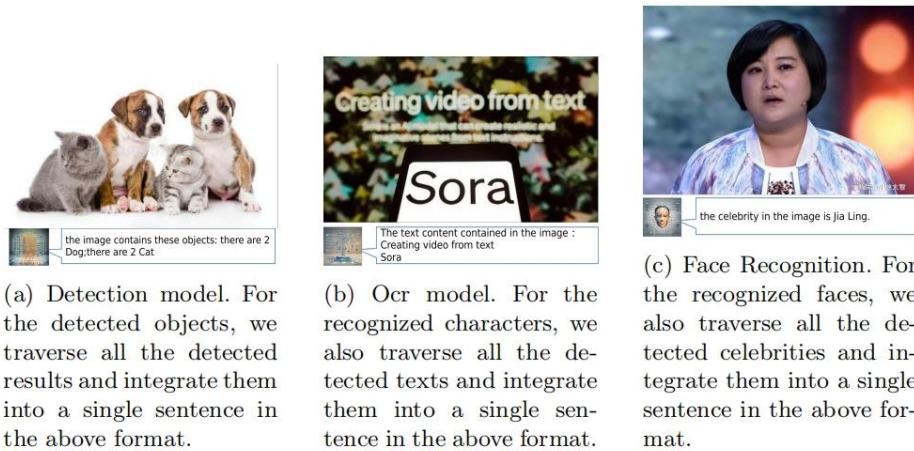


Fig. 2: Details of input formulation. In each sub-image, we adopt a randomly chosen image to exemplify the concrete operation.

#### 3.1 Specialized Models

**Object Detection.** We utilize an object detection model to detect factual information of the input image. To be specific, we adopt PP-YOLOE [24], an industrial state-of-the-art object detector with high performance and friendly deployment, to detect objects inside input image. PP-YOLOE is pre-trained on COCO [25], a large-scale object detection, segmentation, and captioning dataset that has 80 object categories that can cover the most common objects encountered in daily life.

**OCR.** We utilize PaddleOCR<sup>1</sup> to recognize characters inside image. PaddleOCR is an awesome multilingual OCR toolkit based on PaddlePaddle, which supports 80+ language recognition, provides data annotation and synthesis tools, and supports training and deployment among server, mobile, embedded and IoT devices. We utilize this model to extract additional information inside an image

<sup>1</sup> <https://github.com/PaddlePaddle/PaddleOCR>

to serve as supplemental descriptions, together with the detected objects, for the MLLMs to refer to.



Fig. 3: Illustration of our method’s processing. Red words are preprocessed results of specialized models, yellow words are the predefined prompt everybody usually uses, and blue words are the user’s original query, purple words are model’s reply without hallucination. The recognized characters, faces and objects are integrated into one single sentence, which, alongside the user’s original query and image, serves as the final input of MLLMs.

**Face recognition.** We utilize insightface [26] to detect faces inside an image. Insightface is an open-source 2D&3D deep face analysis toolbox, which efficiently implements a rich variety of state-of-the-art algorithms of face recognition, face detection and face alignment, which are optimized for both training and deployment. Furthermore, we establish a repository of celebrities, and the recognized faces are classified as concrete celebrities. These descriptions, alongside the detected objects and PaddleOCR’s characters, also serve as external information for the MLLMs to refer to.

### 3.2 Input Formulation

We utilize the aforementioned specialized traditional deep learning models to detect objects, characters, and faces, and in this part we integrate all these detected

results into a specific format to serve as input, alongside the original user’s question and image, for the MLLMs.

**Object Detection.** For the detected objects, we traverse all the detected results and integrate them into a single sentence in the following format: “*the image contains these objects: there is/are {number} {object}*.”. The detail is exemplified in Fig 2a.

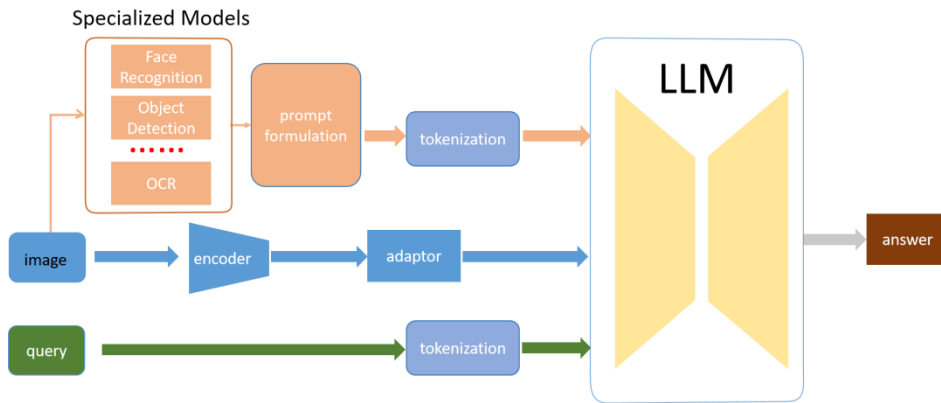


Fig. 4: Flowchart of our method. Given an image and a query, firstly we utilize specialized models to extract descriptions of visual information, these descriptions are then reorganized by prompt formulation block and combined with the original user’s query, the newly combined query and image are then input into the MLLM.

**OCR.** For the recognized characters, we also traverse all the detected texts and integrate them into a single sentence in the following format: “*The text content contained in the image: {recognized characters}*.”. The detail is exemplified in Fig 2b.

**Face recognition.** For the recognized faces, we also traverse all the detected celebrities and integrate them into a single sentence in the following format: “*the celebrity/celebrities in the image is/are: {recognized celebrities}*.”. The detail is exemplified in Fig 2c.

After all these processing, the recognized characters, faces and objects are integrated into one single sentence, which, alongside the user’s original query and image, serves as the final input of MLLMs. The final format of a typical input is exemplified in Fig 3 in detail. In summary, the format is like:

“*Organized OCR results.  
Organized face recognition results.  
Organized detection results.  
Predefined prompt everybody usually uses.  
User’s original query.*”.

Through these steps, our method can directly address the hallucination issue of MLLMs at its root. An overview of our framework is depicted in Fig 4.

Our method utilizes specialized models to generate results, which serve as supplementary descriptions. These reorganized results, alongside the user’s query and image, are then input into the large model, thereby enabling the model to generate correct answers directly by referencing additional factual information.

**Prompt**

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Note that the descriptions may be accompanied by bounding boxes, indicating the position of objects in the image, which are represented as [x1, y1, x2, y2] with floating numbers ranging from 0 to 1. These values correspond to the top left x1, top left y1, bottom right x2, and bottom right y2. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:

- 1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.
- 2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.

Please output a single line for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

[Assistant 1]  
 {Response 1}

[End of Assistant 1]

[Assistant 2]  
 {Response 2}

[End of Assistant 2]

Output format:

Accuracy:  
 Scores of the two answers:  
 Reason:

Detailedness:  
 Scores of the two answers:  
 Reason:

Fig. 5: Prompt template for GPT-4V-aided evaluation. Response 1 and Response 2 are the original responses and the corrected ones, respectively.

Compared to other mainstream methods addressing the hallucination of MLLMs, our method has the following advantages:

- Our method is totally training-free, and requires no re-training of MLLMs, which saves a lot of expenses and time.
- Our framework only requires one inference of one single MLLM.
- Our framework requires no inference of any other large-scale MLLMs, just several traditional deep learning models which are rather small and economical to infer and deploy.



## 4 Experiments

In this section, we will discuss the datasets we use and the experiments we conduct in detail. We use mainstream benchmark datasets POPE [13], MME [14], and LLaVA-QA90 [15], and conduct comprehensive comparative experiments to validate the effectiveness and superiority of our method. Specifically, we choose Qwen-VL-7B and LLaVa-v1.5-13B [7] as our baseline models. Considering that Woodpecker is the most similar training-free method to ours, we also compare our results with theirs utilizing the same baseline model LLaVa-v1.5-13B on POPE and MME benchmarks.

### 4.1 Datasets

**POPE.** The POPE [13] initiative aims to gauge the tendency of MLLMs to produce hallucinations. It employs three varied sampling strategies—random, popular, and adversarial—to construct non-existent object samples. Random sampling randomly selects items not depicted in the image, while popular sampling draws from a pool of frequently seen items not present, and adversarial sampling identifies items often found together but missing from the image. Each kind of strategy has 500 images, and each image has 6 related questions and answers, which is 3000 in total.

For evaluation, to thoroughly compare our method, we directly tested all these images, which amounts to 9,000 in total. The questions balance between positive and negative samples at a 50-50 split. This approach casts object annotations as binary questions, centering on the evaluation of object hallucinations, with a particular emphasis on the aspect of existence. The selected MLLMs will answer like *"Is there a wine glass in the image?"*, and the answer will be measured in a metric of Accuracy, Precision, Recall and F1 Score.

**MME.** The MME [14] is a comprehensive evaluation benchmark for MLLMs. To avoid data leakage that may arise from the direct use of public datasets for evaluation, the annotations of instruction-answer pairs are all manually designed. The concise instruction design can fairly compare MLLMs, instead of struggling in prompt engineering. Besides, with such an instruction, quantitative statistics can also be easily carried out. Also like POPE, The selected MLLMs will also be prompted *Yes or No* questions.

**LLaVA-QA90.** The LLaVA-QA90 [15] contains randomly selected 30 image for COCO-Val-2014, and for each image, three types of questions (conversation, detailed description, complex reasoning) are generated using the proposed data generation pipeline in [15]. Specifically, we sample 10 description-type queries that are paraphrased in various forms to instruct an MLLM to describe an image, such as *"Analyze the image in a comprehensive and detailed manner."* and *"Explain the visual content of the image in great detail."*. GPT-4V [33] is utilized to evaluate the answers generated by the plain baseline model and our framework's model. We directly feed the image to GPT-4V, and prompt it to rate the responses regarding our designed two dimensions, i.e., accuracy and detailedness. The prompt template is available in Fig 5.

## 4.2 Experimental Results

**Results on POPE.** Instead of sampling several hundreds of images, We directly utilized the entire dataset, which amounts to 9,000 image-text queries, thereby enabling a more thorough and comprehensive comparison to demonstrate the superiority of our method. The tested results on POPE are shown in Table 1, which utilizes Qwen-VL-Chat [34] and LLaVa [15] as baseline model. Considering that Woodpecker [12] is the most similar training-free method to

Table 1: Results on POPE using Qwen-VL-7B and LLaVa-v1.5-13B as baseline model. +Piculet denotes MLLM responses generated by our proposed Piculet, and +Woodpecker for woodpecker’s method. The best performances within each setting are bolded. Our method achieves a near-universal advantage across the board.

setting	model	method	Accuracy	Precision	Recall	F1-Score	Yes Rate
adversarial	QWen	plain	0.8417	0.8970	0.772	0.8298	0.4303
		+Piculet	<b>0.8637</b>	<b>0.9225</b>	<b>0.794</b>	<b>0.8535</b>	0.43033
	LLaVA	plain	0.7333	0.6902	<b>0.8467</b>	0.7605	0.6133
		+Woodpecker	0.8067	0.8286	0.7733	0.8000	0.4667
random	QWen	plain	0.8787	0.9805	0.7727	0.8643	0.394
		+Piculet	<b>0.8893</b>	<b>0.9811</b>	<b>0.794</b>	<b>0.8777</b>	0.4047
	LLaVA	plain	0.8600	0.8750	<b>0.8400</b>	0.8571	0.4800
		+Woodpecker	<b>0.8767</b>	0.9593	0.7867	<b>0.8645</b>	0.4100
popular	QWen	plain	0.8657	0.9492	0.7727	0.8519	0.407
		+Piculet	<b>0.8803</b>	<b>0.9597</b>	<b>0.794</b>	<b>0.8690</b>	0.4137
	LLaVA	plain	0.7667	0.7222	<b>0.8667</b>	0.7879	0.6000
		+Woodpecker	0.8067	0.8382	0.7600	0.7972	0.4533
		+Piculet	<b>0.87</b>	<b>0.9641</b>	0.7687	<b>0.8553</b>	0.3987

Table 2: Results on MME using Qwen-VL-7B and LLaVa-v1.5-13B as baseline model. +Piculet denotes MLLM responses generated by our proposed Piculet, and +Woodpecker for woodpecker’s method. The best performances within each setting are bolded. Our method achieves a near-universal advantage across the board.

model	method	Total	Existence	Count	Position	Color	Celebrity	OCR
QWen	plain	871.12	185	140	<b>128.33</b>	180	135.29	<b>102.5</b>
	+Piculet	<b>944.12</b>	<b>190.0</b>	<b>163.34</b>	126.67	<b>185.0</b>	<b>184.12</b>	95.0
LLaVA	plain	698.72	195	95	53.33	78.33	152.06	<b>125.0</b>
	+Woodpecker	-	195	<b>160.00</b>	55.00	155.00	-	-
	+Piculet	<b>928.9</b>	185.0	151.66	<b>121.66</b>	<b>175.0</b>	<b>195.58</b>	110.0

ours, we also compare with their tested results. As can be seen from the results, our proposed framework achieves an across-the-board performance improvement on all test sets and in all aspects. In detail, in all the adversarial, random, and popular testset, our method outperforms both the plain baseline model and the Woodpecker-enhanced model in all the accuracy, precision, recall and f1-score, except the only one random set, where our method is slightly inferior to Wood- pecker.

Table 3: Results on LLaVa-QA90 using Qwen-VL-7B and LLaVa-v1.5-13B as baseline model. with denotes MLLM responses generated by our proposed Piculet. We don't know Woodpecker' exact 10 sampled examples, so cannot compare with their scores. The accuracy and detailedness metrics are on a scale of 10, and a higher score indicates better performance. The best performances within each setting are bolded. Our method achieves better performances on both accuracy and detailedness aspects.

models	method	accuracy	detailedness
QWen	plain	6.1	5.5
	+Piculet	<b>7.3</b>	<b>6.5</b>
LLaVA	plain	5.6	5.9
	+Piculet	<b>6.8</b>	<b>6.3</b>

A seemingly counterintuitive point is that the unenhanced, plain LLaVa actually performs the best on Recall, compared to both Woodpecker and our Piculet. However, this is actually reasonable: because MLLM inherently tends to answer "yes" to all *Yes or No* questions, without discrimination. This results in Recall, a measurement that measures the proportion of correctly classified samples out of all correct samples, being higher than that of both Woodpecker and our Piculet. The relatively high Yes Rate score of plain LLaVA also corroborates this speculation, which attests to our algorithm's effectiveness in mitigating the hallucinations of MLLMs as well.

Overall, our method outperforms Woodpecker, not to mention that our method operates with faster inference and lower resource consumption, while merely pro-

viding additional factual external knowledge to the MLLM, allowing the MLLM to make its own decisions and produce outputs based on the reliable information.

**Results on MME.** The results on MME are shown in Table 2, and the baseline model is also Qwen-VL-Chat and LLaVa. For comparison, we conducted experiments on MME’s existence, count, position, color, celebrity and ocr testset. For comparison, we also add the Woodpecker’s results in the table. As can be seen from the table, our Piculet outperforms Woodpecker in most aspects, with only existence and count set as exceptions, where it slightly lags behind Woodpecker. Even so, considering that our method merely incorporates a few additional small deep learning models apart from the MLLMs to be corrected, and it significantly reduces inference time and operational costs compared to Woodpecker, our method is undoubtedly the better choice.

**Results on LLaVA-QA90.** The results on LLaVA-QA90 is shown in Table 3, and the baseline model is also Qwen-VL-Chat and LLaVa. In this experiment, we sampled 10 description-type queries, which are paraphrased in various forms to instruct an MLLM to describe an image, to evaluate our proposed framework’s performance. The results, as can be seen from the table, show that our method has also achieved superior performance in both evaluation aspects. It’s worth noting that, as Woodpecker’s sampled 10 queries are not exactly known, so we can’t compare with their results here.

### 4.3 Ablation Study.

We conduct an ablation study on MME datasets to validate the superiority and effectiveness of our method. In this section, we utilize Qwen-VL-Chat as baseline model, in each test set, we select two of three specialized models and run experiments to compare the generated results’ scores. The calculated results are shown in Table 4.

Table 4: Ablation study results on MME using Qwen-VL-7B as baseline model. ✓ means results generated utilizing corresponding specialized models.

Detection	OCR	Face	Total	Existence	Count	Position	Color	Celebrity	OCR
	✓	✓	922.40	185.0	140.0	126.67	185.0	183.24	<b>102.5</b>
✓		✓	941.62	190.0	163.33	126.67	<b>190.0</b>	184.12	87.5
✓	✓		888.24	190.0	163.33	126.67	185.0	128.24	95.0
✓	✓	✓	<b>944.12</b>	<b>190.0</b>	<b>163.33</b>	<b>126.67</b>	185.0	<b>184.12</b>	95.0

As can be seen from the experimental results, each specialized model manages to boost the score on its respective test set. Specifically, the results with the specialized Detection model outperform those without it in both existence and countscores. Similarly, the use of the specialized OCR model leads to higher scores on the OCR test set compared to when it is not used. The same can be said for the specialized Face model. Based on the comprehensive comparative

experimental results, we can confidently say that each of our specialized models contributes to improved outcomes. That is, the method we propose, which we name Piculet, can mitigate the hallucination phenomena in MLLMs, making the responses to users' queries more authentic and reliable.

## 5 Conclusion

In this paper, we propose a novel framework named Piculet to address the hallucinations of MLLMs at its root. As a training-free method, our approach requires only single one inference of the target MLLM, and several other small deep-learning models, no other rather large-scale models are involved, which is economical and time-saving, and is plug-and-play in various different MLLMs. We have achieved the goal of reducing hallucinations by supplying the MLLMs with dependable external knowledge generated by specialized models. We evaluate our method on numerous datasets with other methods, and the results demonstrate the effectiveness and improvement of our method. We hope that our method can contribute a small improvement and offer some insights into the handling of hallucinations of MLLMs, thus inspiring further research and development in the field<sup>2</sup>.

## References

1. Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends[J]. *IEEE signal processing magazine*, 2017, 34(6): 96-108.
2. Wang J, Wei Z, Zhang T, et al. Deeply-fused nets[J]. *arXiv preprint arXiv:1605.07716*, 2016.
3. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. *arXiv preprint arXiv:1810.04805*, 2018.
4. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. *Advances in neural information processing systems*, 2020, 33: 1877-1901.
5. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//*International conference on machine learning*. PMLR, 2021: 8748-8763.
6. Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C]//*International Conference on Machine Learning*. PMLR, 2021: 8821-8831.
7. Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning[J]. *arXiv preprint arXiv:2310.03744*, 2023.
8. Wang W, Lv Q, Yu W, et al. Cogvlm: Visual expert for pretrained language models[J]. *arXiv preprint arXiv:2311.03079*, 2023.
9. Zheng K, He X, Wang X E. Minigt-5: Interleaved vision-and-language generation via generative vokens[J]. *arXiv preprint arXiv:2310.02239*, 2023.
10. Gunjal A, Yin J, Bas E. Detecting and preventing hallucinations in large vision language models[J]. *arXiv preprint arXiv:2308.06394*, 2023.

---

<sup>2</sup> The authors have no competing interests to declare that are relevant to the content of this article.

11. Liu F, Lin K, Li L, et al. Mitigating hallucination in large multi-modal models via robust instruction tuning[J]. arXiv preprint arXiv:2306.14565, 2023, 1(2): 9.
12. Yin S, Fu C, Zhao S, et al. Woodpecker: Hallucination correction for multimodal large language models[J]. arXiv preprint arXiv:2310.16045, 2023.
13. Li Y, Du Y, Zhou K, et al. Evaluating object hallucination in large vision-language models[J]. arXiv preprint arXiv:2305.10355, 2023.
14. Fu C, Chen P, Shen Y, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models[J]. arXiv preprint arXiv:2306.13394, 2023.
15. Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. arXiv preprint arXiv:2304.08485, 2023.
16. Zhu D, Chen J, Shen X, et al. Minigt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.
17. Zhang Y, Li Y, Cui L, et al. Siren’s song in the ai ocean: A survey on hallucination in large language models[J]. arXiv preprint arXiv:2309.01219, 2023.
18. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
19. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
20. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
21. Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.
22. Dai W, Liu Z, Ji Z, et al. Plausible may not be faithful: Probing object hallucination in vision-language pre-training[J]. arXiv preprint arXiv:2210.07688, 2022.
23. Lu J, Rao J, Chen K, et al. Evaluation and mitigation of agnosia in multimodal large language models[J]. arXiv preprint arXiv:2309.04041, 2023.
24. Xu S, Wang X, Lv W, et al. PP-YOLOE: An evolved version of YOLO[J]. arXiv preprint arXiv:2203.16250, 2022.
25. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
26. Deng J, Guo J, Ververas E, et al. Retinaface: Single-shot multi-level face localisation in the wild[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5203-5212.
27. Ye Q, Xu H, Xu G, et al. mplug-owl: Modularization empowers large language models with multimodality[J]. arXiv preprint arXiv:2304.14178, 2023.
28. Li B, Zhang Y, Chen L, et al. Otter: A multi-modal model with in-context instruction tuning[J]. arXiv preprint arXiv:2305.03726, 2023.
29. Yin S, Fu C, Zhao S, et al. A Survey on Multimodal Large Language Models[J]. arXiv preprint arXiv:2306.13549, 2023.
30. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
31. Liu S, Zeng Z, Ren T, et al. Grounding dino: Marrying dino with grounded pretraining for open-set object detection[J]. arXiv preprint arXiv:2303.05499, 2023.

32. Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models[J]. arXiv preprint arXiv:2301.12597, 2023.
33. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.
34. Bai J, Bai S, Yang S, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond[J]. 2023.