Yang S. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 9, September - 2024, pg. 62-69.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

OCC-MLLM-Alpha: Empowering Multi-modal Large Language Model for the Understanding of Occluded Objects with Self-Supervised Test-Time Learning

Shuxin Yang[1]*; Xinhan Di[2]

[1]Chinese Academy of Science, Amazon Robotics, Massachusetts, United States
[2]Giant Network AI Lab, Shanghai, China
E-mail: shuxin.y.97@gmail.com

Received date: 9 Sept 2024, Manuscript No. IJCSMA-24-149889 **Editor assigned:** 13 Sept 2024, Pre QC No. IJCSMA-24-149889 (PQ); **Reviewed:** 17 Sept 2024, QC No. IJCSMA-24-149889 (Q); **Revised:** 25 Sept 2024, Manuscript No. IJCSMA-149889 (R); **Published date:** 30 Sept 2024; DOI. 10.5281/zenodo.8116501

Abstract

There is a gap in the understanding of occluded objects in existing large-scale visual language multi-modal models. Current state of the art multi modal models fail to provide satisfactory results in describing occluded objects through universal visual encoders and supervised learning strategies. Therefore, we introduce a multi-modal large language framework and corresponding self-supervised learning strategy with support of 3D generation. We start our experiments comparing with the state of the art models in the evaluation of a large scale dataset SOM Video. The initial results demonstrate the improvement of 16.92% in comparison with the state of the art VLM models.

Keywords: Occlusion handling; Occluded objects understanding; Object recognition; Language model empowerment; Self-supervised learning techniques

1. Introduction

The latest multi-modal dialogue models, such as Mini-Gemini and GPT-4o showed that despite significant progress, their description of large-scale language models for occluded objects remains unsatisfactory. Therefore, we propose OCC-MLLM-Alpha; a visual language model shown in **figure 1** designed to understand occluded objects in image conversations. To achieve this goal, we developed a visual encoder module consisting of the common CLIP model and the proposed 3D model. Additionally, a self-supervised test-time learning strategy with the support of 3D generation is proposed [1].

2. Method

First, we formulate the generative process of the proposed MLLM, named Occlusion-Aware Multimodal Large Language Model (OCC-MLLM-Alpha), for occlusion-aware descriptions of objects at hand. Second, we introduce the formulation details of each proposed OCC-MLLM-Alpha module. Third, the proposed occlusion loss is calculated, and an occlusion-aware training strategy for large multimodal language models is introduced. Fourth, a self-supervised test time training strategy is designed to facilitate the understanding of occluded objects. We represent the generation process of the proposed OCC-MLLM-Alpha into three parts: input formulation, model forwarding, and decoding.

2.1. Formulation of OCC-MLLM-Alpha Generation

2.1.1. Input Formulation: The input of the proposed OCC-MLLM-Alpha consists of images and text. Setting aside specific architectural differences, OCC-MLLM-Alpha generally applies a visual encoder module to extract visual tokens from raw images and uses a cross-modal mapping module to map these tokens to text space as the input of LLM. The mapped visual tokens are used as part of the LLM input along with the text input. The visual tokens are represented as $\{x_v = x_0, x_1, . . . , x_{N-1}\}$. N represents the length of the visual token, which is a fixed number in most cases. Similarly, the input text is tokenized and expressed as $x_p = \{x_n , x_{N+1}, . . . , x_{M+N-1}\}$.The image and text tokens are then concatenated as the final input $\left\{xi\right\}_{t=0}^{T-1}$ where,

$$T=N+M$$

2.1.2. Model Forward: First, OCC-MLLM-Alpha is trained in an auto-regressive manner using causal attention masks, where each token predicts the next token based on the previous token, formally:

$$h = F_{\mathrm{MLLM}^{Occ}}(x_i)$$

$$h = \{h_0, h_1, . . . , h_{T-1}\} \quad (1)$$

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

where, h represents the output hidden states of the last layer of the $F_{MLLM}{}^{Occ}$.

Second, the hidden state h is projected by applying the vocabulary head H via $F_{MLLM}{}^{Occ}$. Get the predicted logits (probability) of the next token, and the calculation is as follows:

$$p\ (x_t \mid x_{<t}) = \mathrm{SoftMax}\ [H(h_t)]_{xt}, x_t \in X \quad (2)$$

where x<t is represented to simplify the sequence $\left\{xi\right\}_{t=0}^{T-1}$ and X is represented as the whole vocabulary set.

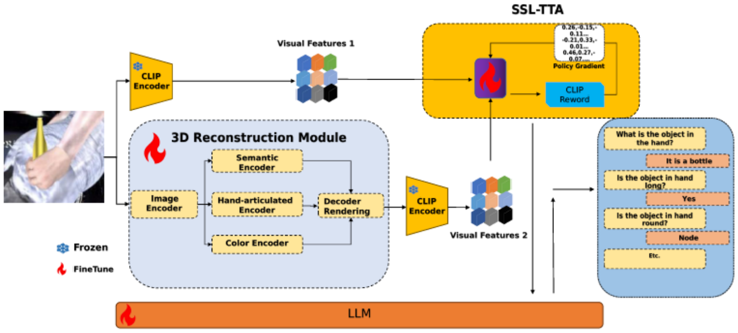


Figure 1. Overview of the proposed multi-modal vision-language model for the occluded objects with self-supervised test-time learning.

2.1.3. Decoding: After applying logits p (x_t | $x_{<t}$), several decoding strategies have been deployed, including greedy decoding, Beam Search, etc. The decoded tokens are concatenated to the last one of the original input text for the next generation round until the end of the generation. The proposed OCCMLLM- Alpha applies a beam search strategy, which is a decoding strategy based on cumulative scores [2].

2.2. Dual Visual Encoder Module

In the forwarding process of the proposed OCC-MLLM Alpha, we designed a new visual encoder module, which consists of two visual encoders. The first visual encoder is the common CLIP, which is used to extract the visual embedding (token) x_v from the RGB input x_{v1} without a specific occlusion representation. The second visual encoder is used to provide a representation of the occluded object visual embedding (token) x_{v2}. Then, the combined representation is calculated as follows:

$$X^v = \alpha \cdot X^{v1} + (1 - \alpha) \cdot X^{v2} \quad (3)$$

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

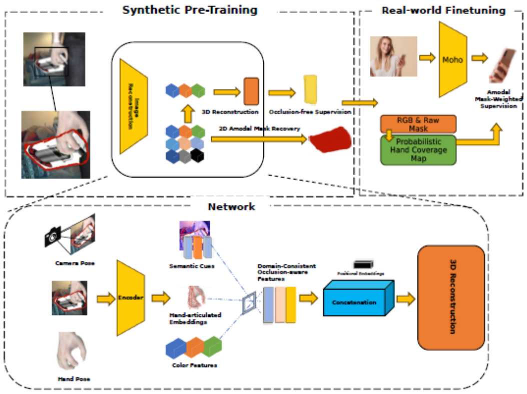


Figure 2. Overview of the proposed second 3D reconstruction module f_{3D}. This method reconstructs a mesh of occluded objects from a single RGB image.

where $\alpha \in [0, 1]$ represents the transparency level of the visual embedding, x_v represents the merged embedding.

2.3. Visual Embedding For Occluded Objects

For the second visual encoder to provide the visual embedding (token) x_{v2} of the occluded object, we designed the second visual encoder f_{3D}, which are composed as follows:
In the first step, the representation of the semantic cues, hand-articulated features and colour features of the occluded object are calculated shown in **figure 1.** These representations are merged into a combination of visual features. The calculation is represented as the following:

$$\begin{aligned} f_{combined} &= f_s\left(f_{cues} + f_{hand} + f_{color}\right) \\ SDF_{object}\left(p\right) &= f_o\left(f_{combined}\right), \end{aligned} \tag{4}$$

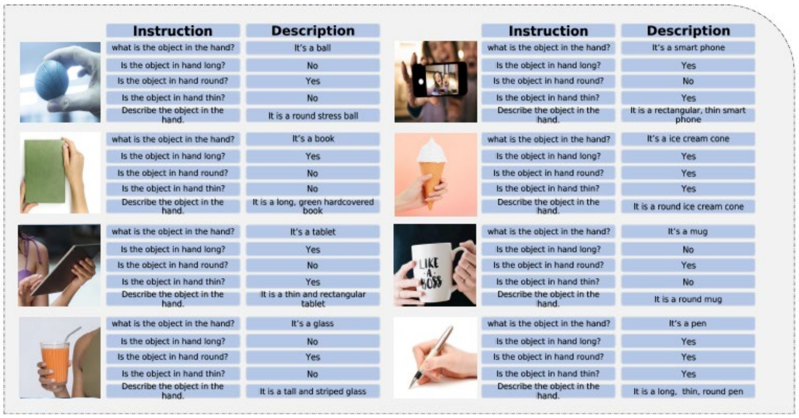
where f_s and f_o are the representation accumulation function and SDF decoder, respectively, p represents the 3D point [3].

In the second step, we apply the calculated SDFs of objects for 3D mesh reconstruction shown in **figure 2.** The computed object $SDF_{object(p)}$ already contains the visual representation of the object under occlusion. We reconstruct the 3D mesh M_{obj} of the occluded object and then project it into the 2D RGB space I_{obj}. To facilitate the use of this 2D visual representation I_{obj} with large language models, we use the visual embedding of X_{v2} as the extracted embedding of the CLIP model [4]. The above calculation is expressed as follows:

$$M_{obj} = f_{recon}(SDF_{object}(p))$$

$$I_{obj} = f_{proj}(M_{obj}) \qquad (5)$$

$$X_{v2} = f_{CLIP}\ (I_{obj})$$



Instruction	Description
what is the object in the hand?	It's a ball
Is the object in hand long?	No
Is the object in hand round?	Yes
Is the object in hand thin?	No
Describe the object in the hand.	It is a round stress ball
what is the object in the hand?	It's a book
Is the object in hand long?	Yes
Is the object in hand round?	No
Is the object in hand thin?	No
Describe the object in the hand.	It is a long, green hardcovered book
what is the object in the hand?	It's a tablet
Is the object in hand long?	Yes
Is the object in hand round?	No
Is the object in hand thin?	Yes
Describe the object in the hand.	It is a thin and rectangular tablet
what is the object in the hand?	It's a glass
Is the object in hand long?	No
Is the object in hand round?	Yes
Is the object in hand thin?	No
Describe the object in the hand.	It is a tall and striped glass

Instruction	Description
what is the object in the hand?	It's a smart phone
Is the object in hand long?	Yes
Is the object in hand round?	No
Is the object in hand thin?	Yes
Describe the object in the hand.	It is a rectangular, thin smart phone
what is the object in the hand?	It's a ice cream cone
Is the object in hand long?	Yes
Is the object in hand round?	Yes
Is the object in hand thin?	Yes
Describe the object in the hand.	It is a round ice cream cone
what is the object in the hand?	It's a mug
Is the object in hand long?	No
Is the object in hand round?	Yes
Is the object in hand thin?	No
Describe the object in the hand.	It is a round mug
what is the object in the hand?	It's a pen
Is the object in hand long?	Yes
Is the object in hand round?	Yes
Is the object in hand thin?	Yes
Describe the object in the hand.	It is a long, thin, round pen

Figure 3. The object is occluded. There are five instructions and five corresponding descriptions.

2.4. Test-Time Adaption Based on Self-Supervised Learning

To enhance the representation of occluded objects for the multi-modal large language model in the test time, we propose a self-supervised learning strategy with the support of 3D generation module. Specifically, a CLIP model is adopted as the reward model and provides feedback for the fine-tuned VLM. Given each test sample, with the support of 3D generation module, the VLM is forced to maximize the CLIP reward between the input and sampled results from the fine-tuned VLM output distribution [5].

The self-supervised training is conducted in the reinforcement learning with rewards. In details, the reward is represented as the following:

$$R\ (t, v) = \text{CLIP}-S\ (t, v) - E_{t\sim P}\ [\text{CLIP}-S\ (t, v)] \qquad (6)$$

Where CLIP−S (t, v) is the self-supervised clip-score on the base of contrastive learning, $E_{t\sim P}$ [CLIP−S(t, v) is the corresponding expectation. Where v is the image and t is the corresponding text.

2.5. Multi-stage Leaning Strategy

At the first stage, the VLM is fine-tuned on the training dataset to perform five specific description tasks **(Figure 3)**. At the second stage, the proposed 3D generation module is trained on the training dataset for 3D reconstruction from a single image. At the third stage, to enhance the representation of the occluded objects, the proposed test-time self-supervised adaption strategy is conducted to force the VLM in the combination with the 3D generation module [6-8].

3. Dataset

We use a large-scale dataset SOM Video containing occluded objects to train the proposed multi-modal large language model to understand them [9].

3.1. Dataset Overview

This dataset SOMVideo consists of a total of 141, 550 scenes with each hand-object scene captured by 10 different views. Each corresponding occlusion-free video clip for supervision is also captured from the same 10 view angles. It also contains 141, 550 × 10 × 5 image-text pairs. This dataset was released to describe occluded objects, and to the best of our knowledge, it is for text descriptions of occluded objects [10]. Besides, we manually calculate the occlusions that about a quarter of the objects are occluded on average, it is important to note that the annotations (text description) of each sample are manually checked. Furthermore, we apply the proposed dataset in the instruction tuning (fine-tuned) stage. All input images are resized to 224 × 224 shown in **figure 3.**

4. Experiments and Results

4.1. Experiments on GPT4o

We first evaluate the performance of GPT4o on the testing portion of the proposed dataset. Four instructions are applied to test each sample in the testing dataset. And the accuracy is demonstrated in the **table 1.** As **table 1** show, the accuracy of the GPT4o is relatively low. In detail, the accuracy for the instruction 1(What's the object in the hand?) is 0.1306, the accuracy for the instruction 2(Is the object in the hand round?) is 0.6910, the accuracy for the instruction 3(Is the object in the hand long?) is 0.6521, the accuracy for the instruction 4(Is the object in the hand thin?) is 0.5839. It demonstrates that GPT4o cannot achieve satisfactory results for the occluded objects [11].

4.2. Experiments on Mini-Gemini

Then, we fine-tuned one epoch for Mini-Gemini using the training set of SOMVideo. The hyper-parameter settings for fine-tuning Mini-Gemini are set as the following: the batch size is 16; the learning rate is 0.00002; the weight attenuation coefficient is 0. As **table 2** shows, in comparison with GPT4o, the accuracy is higher for instruction 1, the accuracy is a little higher for instruction 2, instruction 3 and instruction 4. The visual encoder of the proposed Mini-Gemini is the common clip encoder [9]. As shown in **figure 1** it demonstrates that fine-tuning on a classical multi-modal large language model with a single clip encoder improves the accuracy of the instructions from 0.1306 to 0.4981. However, the accuracy of 0.4981 is still not satisfactory [12].

4.3. Experiments on the Proposed 3D Reconstruction Module

We next explore the capability of the 3D reconstruction module for the test description of the occluded objects. At the stage 1, we train the 3D reconstruction module for the task of 3D reconstruction from a single image. At stage 2, we render the occluded object mesh from the 3D reconstruction module and then project it to 2D RGB space [13]. The rendered RGB image is then described using the fine-tuned VLM for each test image. In the testing phase, we calculate the accuracy of the occluded objects given a single image of the occluded objects. As **table 2** demonstrates, in comparison with the fine-tuned VLM, the accuracy of the instruction 1 for falling testing samples is 0.1692. In detail, there are 6258 occluded samples in the testing set [14-16]. The fine-tuned VLM achieves 4366 correct prediction for the object category classification. Then, the 3D reconstruction module achieves 1128 correct prediction for the left 1892 falling object samples [17].

Table 1. Experimental results of GPT4o and Mini-Gemini.

Model	**GPT4o(Zero-shot)**	**Mini-Gemini**
Instruction 1	0.1306	0.4981
Instruction 2	0.691	0.7284
Instruction 3	0.6521	0.7325
Instruction 4	0.5839	0.7139

Table 2. Accuracy of classification (Instruction 1) for the 3D reconstruction module among falling samples from fine-tuned VLM.

Encoder	**Task**	**Accuracy**
3D Reconstruction [18]	Instruction 1	0.1692

5. Discussion

As the above results demonstrated, the proposed 3D reconstruction module is promising for facilitating the understanding of the occluded objects. We plan to further explore this capability in subsequent experiments. Firstly, the 3D reconstruction module continues to be fine-tuned for the task of the instruction 2, instruction 3 and instruction 4. Secondly, the 3D reconstruction module is merged with the Vision-Language Model (VLM) in a self-supervised learning framework [18].

6. Conclusion

The development of advanced multi-modal dialogue models like Mini-Gemini and GPT-4o has highlighted the ongoing challenges in accurately interpreting occluded objects within image conversations. Although significant advancements have been made, existing large-scale language models still fall short in effectively handling these complexities. In response, we introduced OCC-MLLM-Alpha, a novel visual language model specifically engineered to enhance understanding of occluded objects. Our approach integrates a robust visual encoder module that leverages both the established CLIP model and our innovative 3D model. Furthermore, we propose a self-supervised test-time learning strategy bolstered by 3D generation techniques. Together, these innovations aim to bridge the gap in current capabilities, paving the way for more nuanced and effective interactions in multi-modal environments.

References

1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." *Adv Neural Inf Process Syst* 35 (2022): 23716-23736.
2. Boulanger-Lewandowski, et al. "Audio Chord Recognition with Recurrent Neural Networks." *ISMIR*. 2013.
3. Chen, Gongwei, et al. "Lion: Empowering multimodal large language model with dual-level visual knowledge." *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2024.
4. Chen, Keqin, et al. "Shikra: Unleashing multimodal llm's referential dialogue magic."(2023).
5. Chen, Zerui, et al. "gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction." *Proc. IEEE/CVF Conf Comput Vis Pattern Recognit*. 2023.
6. Gao, Peng, et al. "Llama-adapter v2: Parameter-efficient visual instruction model." (2023).
7. Gong, Tao, et al. "Multimodal-gpt: A vision and language model for dialogue with humans." (2023).
8. Jin, Peng, et al. "Expectation-maximization contrastive learning for compact video-and-language representations." *Adv Neural Inf Process Syst.* 35 (2022): 30291-30306.
9. Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." *Int Conf Mach Learn*. PMLR, 2023.
10. Li, Yanwei, et al. "Mini-gemini: Mining the potential of multi-modality vision language models." (2024).
11. Lin, Bin, et al. "Moe-llava: Mixture of experts for large vision-language models." (2024).
12. Liu, Haotian, et al. "Visual instruction tuning." *Adv Neural Inf Process Syst.* 36 (2024)
13. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *Int Conf Mach Learn., PMLR*, 2021.
14. Wu, Chenfei, et al. "Visual chatgpt: Talking, drawing and editing with visual foundation models." (2023).
15. Yang, Zhengyuan, et al. "Mm-react: Prompting chatgpt for multimodal reasoning and action." (2023).
16. Ye, Qinghao, et al. "mplug-owl: Modularization empowers large language models with multimodality." (2023).
17. Zhang, Chenyangguang, et al. "Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision." *Proc. IEEE/CVF Conf Comput Vis Pattern Recognit*. 2024.
18. Zhu, Deyao, et al. "Minigpt-4: Enhancing vision-language understanding with advanced large language models." (2023).