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

Shuxin Yang
Amazon Robotics
shuxin.y.97@gmail.com

Xinhan Di
Giant Interactive Group Inc AI Lab
dixinhan@ztgame.com

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 SOMVideo [18]. The initial results demonstrate the improvement of 16.92% *in comparison with the state-of-the-art VLM models.*

1. Introduction

The latest multi-modal dialogue models [1, 3, 4, 6–9, 11, 12, 15–17, 19], such as Mini-Gemini [10] and GPT-4o [13] 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 [14] and the proposed 3D model [5]. Additionally, a self-supervised test-time learning strategy with the support of 3D generation is proposed.

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 multi-modal 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 $\mathbf{x}^v = \{x_0, x_1, \ldots, x_{N-1}\}$. 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 $\mathbf{x}^p = \{x_N, x_{N+1}, \ldots, x_{M+N-1}\}$. The image and text tokens are then concatenated as the final input $\{x_i\}_{t=0}^{T-1}$ 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:

$$\begin{aligned} \mathbf{h} &= \mathrm{F}_{\mathrm{MLLM^{Occ}}}(\mathbf{x}_i) \\ \mathbf{h} &= \{h_0, h_1, \ldots, h_{T-1}\} \end{aligned} \tag{1}$$

where $\mathbf{h}$ represents the output hidden states of the last layer of the $\mathrm{F}_{\mathrm{MLLM^{Occ}}}$.

Second, the hidden state h is projected by applying the vocabulary head $\mathcal{H}$ via $\mathrm{F}_{\mathrm{MLLM^{Occ}}}$. Get the predicted log-

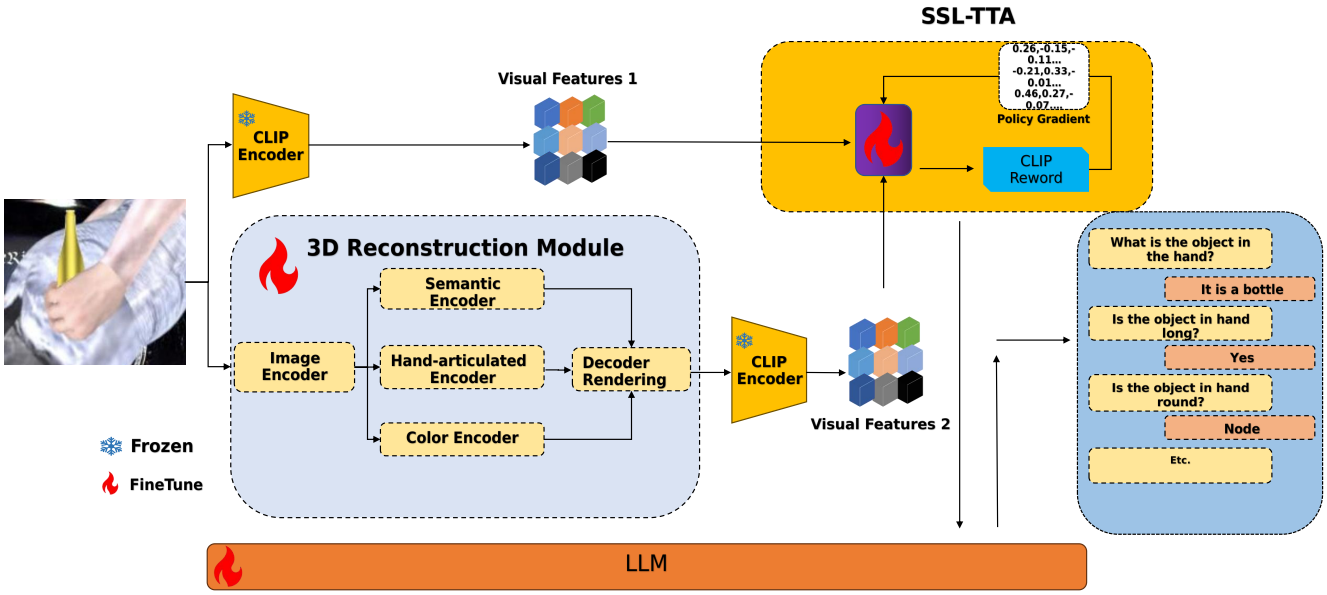

Figure 1. Overview of the Proposed Multi-Modal Vision-Language Model for the Occluded Objects with Self-Supervised Test-Time Learning.

its (probability) of the next token, and the calculation is as follows:

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

where $x_{<t}$ is represented to simplify the sequence $\{x_i\}_{i=0}^{t-1}$ and $\mathcal{X}$ is represented as the whole vocabulary set.

2.1.3 Decoding

After applying logits $p\left(x_t \mid x_{<t}\right)$, several decoding strategies have been deployed, including greedy decoding, Beam Search [2], 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 OCC-MLLM-Alpha applies a beam search strategy [2], which is a decoding strategy based on cumulative scores.

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 [14], which is used to extract the visual embedding (token) x_v from the RGB input $\mathbf{x}_{\text{v1}}$ without a specific occlusion representation. The second visual encoder is used to provide a representation of the occluded object visual embedding(token) $\mathbf{x}_{\text{v2}}$. Then, the combined representation is calculated as follows:

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

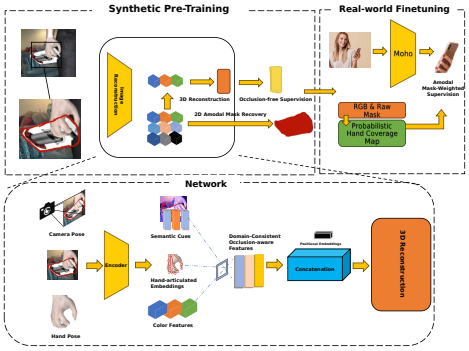

Figure 2. Overview of the proposed second 3D reconstruction module f_{3D}. This method reconstructs a mesh of occluded objects from a single RGB image

where $\alpha \in [0, 1]$ represents the transparency level of the visual embedding, $\mathbf{x}^v$ represents the merged embedding.

2.3. Visual Embedding For Occluded Objects

For the second visual encoder to provide the visual embedding (token) $\mathbf{x}_{\text{v2}}$ of the occluded object, we designed the second visual encoder f_{3D} [10], which is composed as follows:

In the first step, the representation of the semantic cues, hand-articulated features and color features [10] of the occluded object are calculated (shown in Figure 2). These representations are merged into a combination of visual fea-

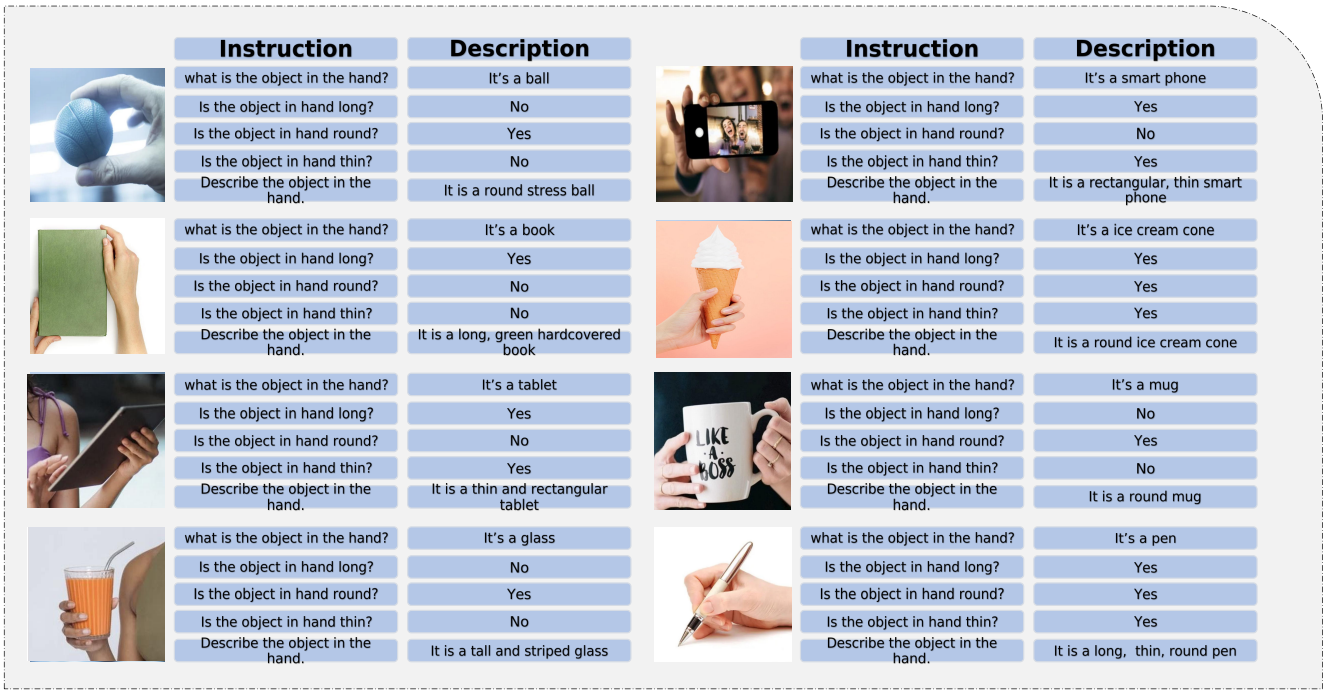

Figure 3. Dataset example. The object is occluded. There are five instructions and five corresponding descriptions.

tures. The calculation is represented as the following:

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

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

In the second step, we apply the calculated SDFs of objects for 3D mesh reconstruction (shown in Figure 2). The computed object $\mathrm{SDF}_{\mathrm{object}}(p)$ 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 $\mathbf{x}_{\mathrm{v2}}$ as the extracted embedding of the CLIP model [14]. The above calculation is expressed as follows:

$$\begin{aligned} M_{obj} &= f_{recon}(\mathrm{SDF}_{\mathrm{object}}(p)) \\ I_{obj} &= f_{proj}(M_{obj}) \\ \mathbf{x}_{\mathrm{v2}} &= f_{CLIP}(I_{obj}) \end{aligned} \tag{5}$$

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 [14] is adopted as the reward model and provides feedback for the fine-tuned VLM [10]. Given each test sample, with the support of 3D generation module [18], the VLM [10] is forced to maximize the CLIP [14] reward between the input and sampled results from the fine-tuned VLM [10] output distribution.

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

$$\mathcal{R}(\boldsymbol{t}, \boldsymbol{v}) = \mathrm{CLIP-S}(\boldsymbol{t}, \boldsymbol{v}) - \mathbb{E}_{\boldsymbol{t} \sim P}[\mathrm{CLIP-S}(\boldsymbol{t}, \boldsymbol{v})] \tag{6}$$

Where $\mathrm{CLIP-S}(\boldsymbol{t}, \boldsymbol{v})$ is the self-supervised clip-score on the base of contrastive learning [14], $\mathbb{E}_{\boldsymbol{t} \sim P}[\mathrm{CLIP-S}(\boldsymbol{t}, \boldsymbol{v})$ 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 [10] is fine-tuned on the training dataset [18] to perform five specific description tasks (Figure 3). At the second stage, the proposed 3D generation module is trained on the training dataset [18] 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 [10] in the combination with the 3D generation module [18].

3. Dataset

We use a large-scale dataset SOMVideo [18] containing occluded objects to train the proposed multi-modal large language model to understand them.

3.1. Dataset Overview

This dataset SOMVideo [18] 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 $141,550 \times 10 \times 5$ 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. 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 [13]

We first evaluate the performance of GPT4o [13] 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 shows, the accuracy of the GPT4o [13] 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 [13] cannot achieve satisfactory results for the occluded objects.

4.2. Experiments on Mini-Gemini [10]

Then, we fine-tuned one epoch for Mini-Gemini [10] using the training set of SOMVideo [18]. The hyper-parameter settings for fine-tuning Mini-Gemini [10] 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 [13], 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 [10] is the common clip encoder[14]. (Shown in Figure 1). It demonstrates that fine-tuning on a classical multi-modal large language model [10] with a single clip encoder [14] improves the accuracy of the instructions from 0.1306 to 0.4981. However, the accuracy of 0.4981 is still not satisfactory.

4.3. Experiments on the Proposed 3D Reconstruction Module [18]

We next explore the capability of the 3D reconstruction module [18] for the test description of the occluded objects. At the stage 1, we train the 3D reconstruction module [18] 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. The rendered RGB image is then described using the fine-tuned VLM [10] 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 [10], the accuracy of the instruction 1 for falling testing samples [10] is 0.1692. In detail, there are 6258 occluded samples in the testing set [18], the fine-tuned VLM [10] achieves 4366 correct prediction for the object category classification. Then, the 3D reconstruction module [18] achieves 1128 correct prediction for the left 1892 falling object samples.

Table 1. Experimental results of GPT4o and Mini-Gemini

Model	GPT4o(Zero-shot)	Mini-Gemini
Instruction 1	0.1306	0.4981
Instruction 2	0.6910	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 [18]

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

4.4. Future Experiments

As the above results demonstrated, the proposed 3D reconstruction module [18] 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 [18] continues to be fine-tuned for the task of the instruction 2, instruction 3 and instruction 4. Secondly, the 3D reconstruction module [18] is merged with the Vision-Language Model(VLM) [10] in a self-supervised learning framework.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch,

Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[2] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Audio chord recognition with recurrent neural networks. In *ISMIR*, pages 335–340. Curitiba, 2013. 2

[3] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. *arXiv preprint arXiv:2311.11860*, 2023. 1

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1

[5] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction, 2023. 1

[6] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1

[7] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

[8] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in Neural Information Processing Systems*, 35:30291–30306, 2022.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[10] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*, 2023. 1, 2, 3, 4

[11] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 1

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[13] OpenAI. Hello gpt-4o. 1, 4

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3, 4

[15] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 1

[16] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

[17] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1

[18] Chenyangguang Zhang, Guanlong Jiao, Yan Di, Gu Wang, Ziqin Huang, Ruida Zhang, Fabian Manhardt, Bowen Fu, Federico Tombari, and Xiangyang Ji. Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9992–10002, 2024. 1, 3, 4

[19] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1