Unveiling the Invisible: Reasoning Complex Occlusions Amodally with AURA

Zhixuan Li[1] Hyunse Yoon[2] Sanghoon Lee[2] Weisi Lin[1*]
[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]Department of Electrical and Electronic Engineering, Yonsei University, Korea
zhixuanli520@gmail.com, hsyoon97, slee@yonsei.ac.kr, wslin@ntu.edu.sg

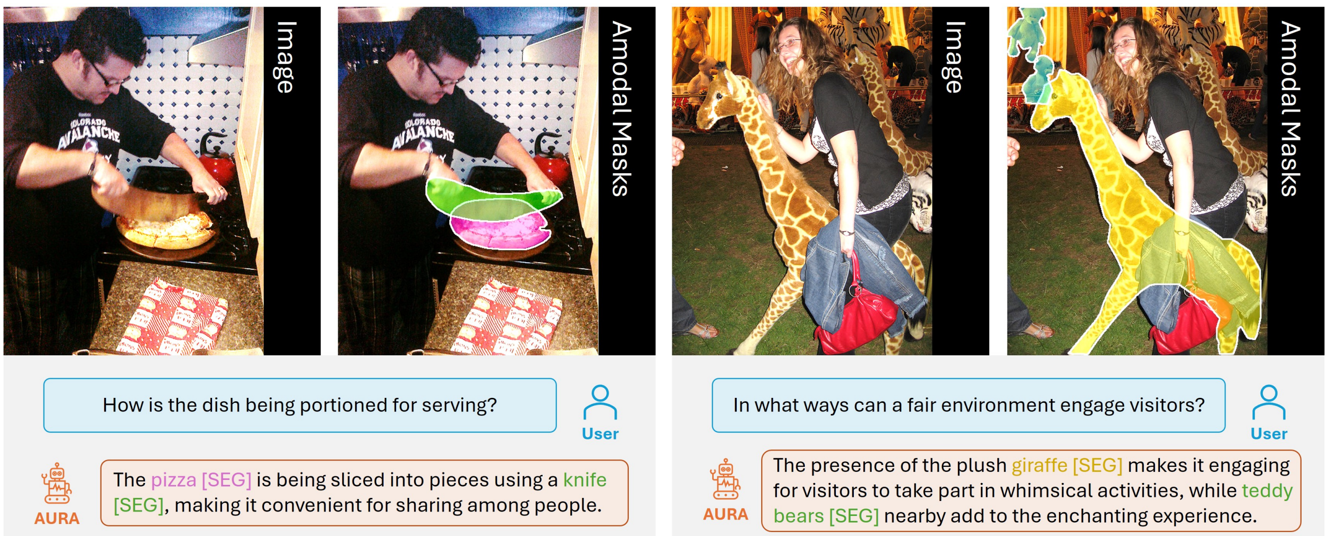

Figure 1. We introduce AURA, a multi-modal approach designed for reasoning the amodal segmentation mask including both visible and occlusion regions based on the user's question. AURA can deduce the implicit purpose underneath the question and respond with the textual answer along with predicted amodal masks for various objects. Best viewed in color.

Abstract

Amodal segmentation aims to infer the complete shape of occluded objects, even when the occluded region's appearance is unavailable. However, current amodal segmentation methods lack the capability to interact with users through text input and struggle to understand or reason about implicit and complex purposes. While methods like LISA integrate multi-modal large language models (LLMs) with segmentation for reasoning tasks, they are limited to predicting only visible object regions and face challenges in handling complex occlusion scenarios. To address these limitations, we propose a novel task named amodal reasoning segmentation, aiming to predict the complete amodal shape of occluded objects while providing answers with elaborations based on user text input. We develop a generalizable dataset generation pipeline and introduce a new dataset focusing on daily life scenarios, encompassing diverse real-world occlusions. Furthermore, we present AURA (Amodal Understanding and Reasoning Assistant), a novel model with advanced global and spatial-level designs specifically tailored to handle complex occlusions. Extensive experiments validate AURA's effectiveness on the proposed dataset. The code, model, and dataset are released in this page.

*Corresponding author.

1. Introduction

With the development of the Segment Anything Model (SAM) series methods [24, 25, 58], predicting the segmentation mask of visible regions has been largely solved. Amodal segmentation aims to predict the entire shape of an object, including both visible and occluded regions, which is naturally challenging since visual cues for the occluded regions are unavailable. Predicting the intact shape of occluded objects is crucial for various real-world applications, such as autonomous driving [12, 45, 50], robotics [2, 5, 22], and image editing [59, 63].

Existing amodal segmentation approaches have covered the fields including semantic [3, 49], instance [28, 37, 71], and panoptic [42] segmentation. Although existing amodal segmentation methods [10, 14, 45, 57] can address the problem of inferring occluded object shapes to some extent, they lack the ability to integrate with large language models (LLMs) [53, 54, 65], especially in performing reasoning and segmentation based on scene context, limiting their ability in interacting with users and understanding the implicit purpose underneath users questions. Meanwhile, although existing reasoning segmentation methods [26, 55] can work well on predicting the visible segmentation, they cannot predict the complete amodal segmentation mask, limiting their ability to reason in scenarios with complex occlusions. As a matter of fact, there is currently a significant gap in the availability of datasets designed for amodal reasoning segmentation, as well as in the development of algorithms specifically tailored to this task.

To address this gap, we propose a novel task named Amodal Reasoning Segmentation (ARS). As shown in Figure 1, the ARS task aims to interact with the user through textual questions, reasoning in the complex scene, and predicting accurate segmentation masks with textual answers. For this task, we present a highly generalizable data collection pipeline and construct a new dataset, AmodalReasonSeg, which covers daily life scenes (encompassing both indoor and outdoor environments) derived from the challenging COCOA-cls dataset [10, 71]. The AmodalReasonSeg dataset facilitates object reasoning and segmentation for visible and amodal masks across diverse scenes.

Along with the new task and dataset, we introduce a novel model named AURA (**A**modal **U**nderstanding and **R**easoning **A**ssistant). AURA can understand the user's input questions, reason about the scene in the image, and predict both visible and amodal segmentation masks simultaneously. Given the challenge of predicting the shape of occluded parts, which are invisible in the image, we propose two key designs to enhance occlusion prediction capabilities. Specifically, we introduce the Occlusion Condition Encoder, which can distinguish whether objects are occluded and assess the occlusion degree comprehensively. Besides, the Spatial Occlusion Encoder can provide spatial guidance on whether each region is occluded, ensuring consistency between visible and amodal mask predictions. It is worth noticing that the proposed AmodalReasonSeg dataset and the designed AURA method support reasoning and mask prediction for various objects within a single-turn conversation, enhancing applicability to real-world scenarios. Furthermore, AURA supports multi-round conversations, increasing its versatility.

Extensive experiments on the proposed AmodalReasonSeg dataset show that AURA outperforms existing methods in prediction accuracy, reasoning capability, and generalizability for complex occlusion scenarios.

The contributions of this paper can be summarized as follows: (1) We propose a novel task: amodal reasoning segmentation, along with a generalizable dataset collection pipeline. We construct a new dataset, AmodalReasonSeg, covering a variety of complex scenes from daily life. (2) We introduce AURA, a multi-modal method integrating vision and language for complex reasoning, capable of predicting both visible and amodal segmentation masks accurately. To the best of our knowledge, AURA is the first method explicitly designed for reasoning segmentation of amodal regions. (3) Extensive experiments demonstrate the effectiveness of AURA through both quantitative and qualitative evaluations.

2. Related Work

2.1. Amodal Segmentation

Amodal segmentation [28] aims to predict the complete shape of an occluded object, encompassing both visible and occluded parts. Several methods [10, 28, 30, 32, 37, 38, 45] are developed by extending the methods designed for visible segmentation or object detection [4, 18–20, 23, 64], while the relative occlusion order information is leveraged by many methods [43, 63, 66, 68]. Since shape prior knowledge is an essential approach for inferring the shape of the occluded region, several methods [6, 9, 11, 29, 31, 34, 52, 57] are designed to learn and utilize it. For instance, VRSP [57] designs a codebook to store and retrieve shape priors for amodal masks. A3D [29] learns to reconstruct 3D shape priors to guide amodal segmentation. C2F-Seg [11] works in a coarse-to-fine manner equipped with a Vector-Quantized codebook [9]. Moreover, various datasets have been proposed for diverse scenarios, including daily life [10, 71], indoor [7, 8, 62, 70], autonomous driving [42, 45, 47], industrial [10], shopping [32], and gaming [14]. However, existing amodal segmentation methods and datasets are limited to using images or videos as input, focusing solely on predicting amodal masks without user interaction or reasoning about implicit user purposes. In this paper, we present the first high-quality dataset for the newly proposed amodal reasoning segmentation task and introduce a Vision-Language Model-based approach specifically designed for this task.

2.2. Large Models for Reasoning Segmentation

With the rapid advancement of Large Language Models (LLMs) [53, 54, 65] such as GPT-4 [1] and LLaMA [53], language-based approaches have demonstrated impressive zero-shot generalization abilities across diverse real-world applications. Additionally, Vision Language Models

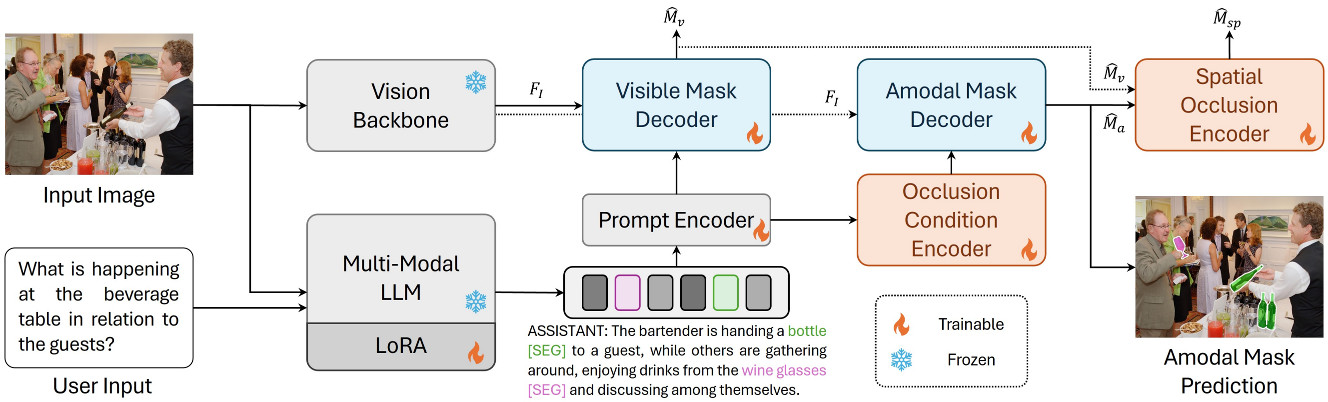

Figure 2. Overall architecture of the proposed AURA. (a) Given an input image and the input question from the user, the Vision Backbone extracts the visual features of the input image, and the Multi-Modal LLM equipped with LoRA is utilized for understanding the input image and textual questions simultaneously and responding with textual answers that include the [SEG] tokens indicating the segmentation masks. (b) For each [SEG], the Prompt Encoder takes its embedding of the Multi-Modal LLM and outputs the refined embeddings corresponding to the [SEG]. (c) Finally, the Visible Mask Decoder predicts the visible mask using each [SEG]'s refined embeddings. The Amodal Decoder predicts the amodal mask using the occlusion-aware embedding predicted by the Occlusion Condition Encoder. A Spatial Occlusion Encoder is designed to constrain the spatial occlusion information of the predicted visible and amodal segmentation masks to be accurate. Best viewed in color.

(VLMs) [21, 35, 36] like BLIP-2 [27] and LLaVA [36] have achieved state-of-the-art performance in comprehensively understanding both visual and textual information. To leverage the reasoning capabilities of VLMs, LISA [26] introduces the task of reasoning segmentation, which predicts visible segmentation masks by reasoning based on user input questions. Building upon LISA, several recent methods, such as Pixel-LM [48], VISA [60], and LLM-Seg [55], have been proposed to enhance different aspects of the reasoning segmentation task, such as multi-object segmentation, video segmentation, and training efficiency, respectively. Despite these advancements, all existing methods are limited to segmenting only the visible region of target objects, leaving segmenting and analyzing the occluded region untouched. Our work aims to bridge this gap by enabling reasoning and segmenting in an amodal manner, specifically designed to handle complex occlusions in real-world scenarios.

3. Methodology

In this section, we first introduce the definition of the proposed amodal reasoning segmentation task in 3.1. Then the overall architecture of the designed method is presented in 3.2. Next, the special designs for learning and handling occlusion are introduced in 3.3. Finally, the loss functions that organize the training objectives are shown in 3.4.

3.1. Task Definition

For the proposed amodal reasoning segmentation (ARS) task, each image x_{img} of the dataset is associated with various questions x_{text}, answers y_{text}, and segmentations y_{seg}. For each pair, x_{text} is a question from the user input that does not directly point to objects in the image. y_{text} is the ground-truth textual answer that responds to the question and includes the description of one or more objects alongside the [SEG] token representing the position of the segmentation mask. y_{seg} represents the ground-truth visible and amodal segmentation masks (M_v and M_a), corresponding to the [SEG] in the answer y_{text}.

3.2. Overall Architecture

As shown in Figure 2, the overall architecture of the proposed AURA includes three stages. Specifically, (1) the Vision Backbone takes the input image I as input for extracting its visual feature F_I, and the Multi-Modal LLM $\mathcal{E}_{MLLM}$ equipped with LoRA is employed to understand the input image I and the user's input question x_{text}, and predict the textual answer $\hat{y}_{text}$ to respond the question. (2) The Prompt Encoder $\mathcal{E}_P$ takes the embedding e_{mllm} of each [SEG] from the hidden states of the Multi-Modal LLM and predicts the refined embedding e_r for each [SEG]. (3) Finally the e_r and F_I are fed into the Visible Mask Decoder $\mathcal{D}_V$ for predicting the visible segmentation mask $\hat{M}_v$ of each [SEG]. Besides, the e_r is used by the Occlusion Condition Encoder $\mathcal{E}_{OC}$, which learns to predict the occlusion-aware embedding e_{oa} that is contained within the occlusion condition information. Then the Amodal Mask Decoder $\mathcal{D}_A$ uses the F_I and e_{oa} for predicting the amodal segmentation mask $\hat{M}_a$ of each [SEG]. The Spatial Occlusion Encoder $\mathcal{E}_{SO}$ is designed to constrain the predicted $\hat{M}_v$ and $\hat{M}_a$ to conform to the spatial occlusion relationship, which implic-

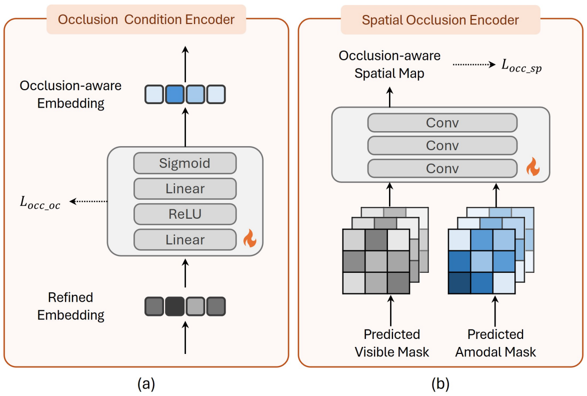

Figure 3. The designed occlusion encoders. (a) The Occlusion Condition Encoder learns to embed the occlusion condition of each mask and output the occlusion-aware embedding. (b) The Spatial Occlusion Encoder takes the predicted visible and amodal masks as input and predicts the occlusion-aware spatial map to constrain the spatial relationship between the two predicted masks.

itly improves the accuracy of the two predicted masks.

The textual prediction of the Multi-Modal LLM is supervised using the $\mathcal{L}_{text}$, which can be denoted as:

$$\mathcal{L}_{text} = \mathcal{L}_{CE}(\hat{y}_{text}, y_{text}), \tag{1}$$

where $\mathcal{L}_{CE}$ is the Cross-Entropy loss [61]. Moreover, for supervising the visible and amodal mask predictions, the $\mathcal{L}_{mask}$ is used as shown in the following:

$$\begin{aligned} \mathcal{L}_{mask} = \mathcal{L}_{CE}(\hat{M}_v, M_v) + \mathcal{L}_{CE}(\hat{M}_a, M_a) + \\ \mathcal{L}_{Dice}(\hat{M}_v, M_v) + \mathcal{L}_{Dice}(\hat{M}_a, M_a), \end{aligned} \tag{2}$$

where $\mathcal{L}_{Dice}$ is the Dice loss [51].

3.3. Occlusion Learning

Figure 3 illustrates the designed occlusion encoders for handling occlusion learning in AURA. Detailed network structures, inputs and outputs, and losses are presented.

Occlusion Condition Encoder is designed to improve the ability of the refined embedding e_r predicted by the Prompt Encoder to be aware of the occlusion condition of the object(s) corresponding to the [SEG]. Specifically, the e_r is fed into the Occlusion Condition Encoder $\mathcal{E}_{OC}$ and predicts the occlusion-aware embedding e_{oa} that is distinguishable by the degree of occlusion. The Occlusion Condition Encoder $\mathcal{E}_{OC}$ additionally predicts the occlusion rate $\hat{r}$ of the object, which reflects the occlusion condition comprehensively and is supervised by $\mathcal{L}_{occ-oc}$ with the ground-truth occlusion rate r. Moreover, a Linear layer and a Sigmoid layer are employed to predict the occlusion rate $\hat{r}$ based on the refined embedding e_r.

Spatial Occlusion Encoder is proposed to constrain the spatial occlusion conditions between the predicted visible masks $\hat{M}_v$ and amodal masks $\hat{M}_a$, enhancing the accuracy of visible and amodal segmentation prediction. Specifically, the Spatial Occlusion Encoder $\mathcal{E}_{SO}$ takes both $\hat{M}_v$ and $\hat{M}_a$ as input and predicts the occlusion-aware spatial map $\hat{M}_{sp}$, which is supervised by $\mathcal{L}_{occ-sp}$ with the ground-truth occlusion-aware spatial map M_{sp}. The ground truth M_{sp} is defined by setting regions belonging to the ground-truth visible mask M_v as 1 and setting regions belonging to the ground-truth amodal mask M_a but not M_v as 2, and setting other background regions as 0. In this approach, the constraint between predicted masks $\hat{M}_v$ and $\hat{M}_a$ is constructed and contributes to accurate segmentation.

For supervising the predictions of the occlusion condition and the spatial occlusion map, the $\mathcal{L}_{occ}$ is used as shown below:

$$\mathcal{L}_{occ} = \mathcal{L}_{MSE}(\hat{r}, r) + \mathcal{L}_{CE}(\hat{M}_{sp}, M_{sp}), \tag{3}$$

where $\mathcal{L}_{MSE}$ is the Mean Squared Error loss.

3.4. Loss Functions

The overall loss function $\mathcal{L}$ consists of four parts, supervising the textual answers $\hat{y}_{text}$ predicted by the Multi-Modal LLM, the predicted visible and amodal masks ($\hat{M}_v$ and $\hat{M}_a$), and the predictions of two designed occlusion encoders including the occlusion rate $\hat{r}$ and the occlusion-aware spatial map $\hat{M}_{sp}$. The overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{text} + \mathcal{L}_{mask} + \mathcal{L}_{occ}. \tag{4}$$

4. AmodalReasonSeg Dataset

Existing amodal segmentation datasets [10, 45, 62] only contain images/videos and ground-truth masks without language annotations, which cannot fulfill the requirement of the amodal reasoning segmentation task. To this end, it is essential to construct a large-scale dataset with high-quality language annotations, including implicit user questions and the corresponding answers.

In this section, we first introduce the basic information of the newly constructed dataset in 4.1. Next, we demonstrate a general dataset generation pipeline that is generable for diverse scenarios and can be applied to generate new datasets efficiently in 4.2.

4.1. Overview of the Dataset

To construct the new dataset for the amodal reasoning segmentation task, the challenging amodal segmentation dataset COCOA-cls [10, 71] is utilized. The COCOA-cls dataset is built upon the famous COCO [33] dataset, which includes daily scenarios with indoor and outdoor scenes.

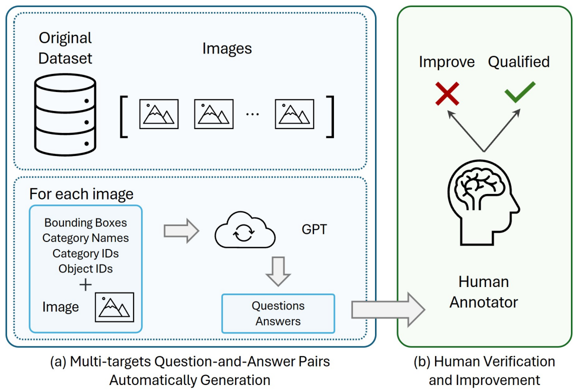

Figure 4. The designed generalizable pipeline for dataset generation. (a) Each image and the information of objects in it are gathered and fed into the GPT to generate questions and answers. (b) Then human annotators will check the generated data and improve it manually. Cross-check ensures the quality of the generated data.

For benchmarking the proposed amodal reasoning segmentation (ARS) task, we present the AmodalReasonSeg dataset, which is built upon the COCOA-cls dataset. The AmodalReasonSeg dataset includes 3,143 images and 35,494 high-quality question-and-answer pairs corresponding to the visible and amodal segmentation masks. The visible and amodal bounding boxes and the category information are included in our dataset. For each image, there are 11.3 pairs of questions and answers on average annotated to cover diverse potential conversations. Overall, there are 2,018 images and 22,856 pairs of questions and answers in the train set and 1,125 images and 12,638 question-and-answer pairs in the test set.

Compared with existing amodal segmentation datasets, the AmodalReasonSeg dataset supports interacting with users by using textual questions as input, and the answers respond to the implicit purpose underneath the questions and alongside the visible and amodal segmentation predictions. Moreover, it is worth noticing that the answer corresponding to each image could refer to various objects, meaning the proposed dataset supports reasoning and predicting different targets in a single conversation, which is essential in real-world applications.

4.2. Dataset Generation

A generalizable pipeline is designed to generate the new dataset, as shown in Figure 4. Since manually annotating the question-and-answer pairs and the visible and amodal segmentation masks is labor-intensive and time-consuming, we present a semi-automatic pipeline inspired by [48] to keep the efficiency of the annotating process and high-quality annotations. Specifically, the images and annotations, including category information and visible and amodal masks, are reserved from the original dataset. For generating the question-and-answer pairs, each image alongside the annotations is organized by a designed prompt template to be understood by a powerful and versatile large Vision-Language model, which is the ChatGPT-4o in our work that could take both image and text as input and output text response.

The prompt template is carefully designed to require ChatGPT-4o to generate question-and-answer pairs that satisfy the requirements to ensure the generation of high-quality data. The annotation information of objects for this image is collected and fed to ChatGPT-4o. For each image, ten question-and-answer pairs are required to be generated. Different objects are required to be organized and contained in one answer. Besides, the relative occlusion information between different objects is required to be considered when generating the question-and-answer pairs, which makes the text description in the answer helpful for discriminating the occlusion relationship between objects. The existing annotations from the original dataset of all objects in the image are organized and taken as input for the ChatGPT-4o. Finally, three human annotators check the generated questions and verify if they are qualified. During human verification, qualified samples are retained, while flawed ones are revised with appropriate questions and answers instead of being discarded. Common issues include incorrect object references, implausible functions, or inaccurate quantities. For the ones that need to be improved, human annotators will manually design appropriate questions and answers for these images. Cross-check is utilized between different human annotators to ensure the quality of the generated data.

5. Experiments

5.1. Experimental Settings

Dataset. All experiments, including our proposed method and compared methods, are trained on the training set of the proposed AmodalReasonSeg dataset, and evaluated on the validation set of this dataset for fairness. No extra dataset besides the AmodalReasonSeg dataset is used.

Model Structure. We employ the LLaVA-7B-v1-1 [36] as the multi-modal LLM, which is equipped with LoRA for efficient fine-tuning. For the Vision Backbone and the Prompt Encoder, the SAM [25] method with the ViT-H version is utilized. The mask decoder of SAM is employed for the Visible Mask Decoder and Amodal Mask Decoder. Unless specifically specified, all experiments of the proposed AURA method are based on the model structure of using the LLaVA-7B-v1-1 [36] as the multi-modal LLM and the ViT-H version of SAM [25].

Implementation Details. We implement our proposed AURA based on Pytorch [44]. The AdamW [39] optimizer is used with the learning rate set to 1e-3. The WarmupDecayLR learning rate adjustment policy is used with 100 iter-

Method	Amodal		Visible	
	gIoU	cIoU	gIoU	cIoU
Amodal Segmentation Methods				
VRSP† [57] AAAI'21	15.18	16.55	17.41	19.75
C2F-Seg† [11] ICCV'23	18.62	18.25	19.12	20.31
SD-Amodal† [62] CVPR'24	20.87	20.09	21.52	23.10
MLLM-based Methods				
LISA† [26] CVPR'24	23.05	22.87	24.39	26.81
LLM-Seg† [55] CVPR'24	24.62	23.06	25.19	27.94
GSVA† [56] CVPR'24	42.30	42.87	47.01	48.22
OMG-LLaVA† [67] NeurIPS'24	43.92	44.21	49.41	49.38
GLaMM† [46] CVPR'24	45.75	46.82	49.93	54.14
PSALM† [69] ECCV'24	<u>46.81</u>	<u>47.03</u>	<u>50.64</u>	<u>55.19</u>
Baseline	43.55	43.92	48.96	48.63
AURA	**47.76**	**47.32**	**51.31**	**55.38**

Table 1. Comparison of the amodal reasoning segmentation performance on the AmodalReasonSeg dataset. † denotes that this method is trained twice with ground-truth visible and amodal masks of the proposed AmodalReasonSeg dataset, respectively. Best in **bold**, second <u>underlined</u>.

ations for warmup and then linearly decreases the learning rate to 0. Apart from the ablation studies using 13B models as Multi-Modal LLM Backbone that use two NVIDIA A100 GPUs (80GB Memory) for experiments, all of our experiments are conducted on two NVIDIA V100 GPUs (32GB Memory). The batch size for each GPU card is set to 1. The entire training iterations are 5000, and the gradient accumulation technique is used by setting the accumulation step to 4. The entire training time is about 10 hours.

Evaluation Metrics. The gIoU and cIoU metrics are used for evaluating the segmentation performance of amodal and visible masks. The gIoU is calculated by averaging the Intersection-over-Unions (IoUs) of all images, while the cIoU is calculated by dividing the cumulative intersection by the cumulative union of all masks.

5.2. Amodal Reasoning Segmentation

Compared Methods. To benchmark the performance of amodal reasoning segmentation on the proposed AmodalReasonSeg dataset, state-of-the-art amodal segmentation methods [11, 57, 62] and multimodal large language model (MLLM)-based methods [26, 46, 55, 56, 67, 69] for reasoning segmentation of visible region are used for comparison. All methods are trained and evaluated on the train set and the val set of our proposed AmodalReasonSeg dataset, ensuring fairness. LLaVA-7B-v1-1 is employed for all MLLM-based methods.

The amodal segmentation methods [11, 57, 62] are designed for taking an image as input and predicting amodal segmentation masks for all objects, but cannot understand the user's question, reasoning, and response with textual answers and segmentations. Besides, MLLM-based methods [26, 46, 55, 56, 67, 69] are designed to segment the visible regions of each object by taking the image and the user's question as input.

However, these aforementioned methods can only predict either visible or amodal masks by training twice with ground-truth visible or amodal masks, while the proposed AURA can reason and predict the visible and amodal masks simultaneously. Moreover, MLLM-based methods like LISA can only reason for one object to each question, while the proposed AURA could reason and respond with various objects with textual answers describing them and corresponding segmentation predictions. The baseline method is the proposed AURA without using the designed two occlusion learning modules, including the Occlusion Condition Encoder $\mathcal{E}_{OC}$ and the Spatial Occlusion Encoder $\mathcal{E}_{SO}$.

Quantitative Results. Table 1 shows the performance of all compared methods on the proposed AmodalReasonSeg dataset. † means this method is trained twice on ground-truth visible and amodal masks, respectively, due to a lack of the ability to predict both masks at the same time. AURA achieves the best performance among all metrics regarding amodal and visible segmentations. To be more specific: (1) Firstly, since amodal segmentation methods [11, 57, 62] are designed for amodal segmentation and cannot take a textual question as input and understand it, they cannot be certain of which objects to segment, and the predicted masks could not be for the target objects. Therefore, these methods are evaluated by computing metrics between all predicted objects and target objects. In contrast, the proposed AURA can understand the purpose underneath the user's question and predict the amodal and visible masks for the targeted objects mentioned in the generated textual answer. (2) Secondly, since these MLLM-based methods are designed for segmenting visible regions of objects, their performance is not as well as AURA, which is specifically designed for the amodal reasoning segmentation task with task-specific designs for handling occlusions. Moreover, methods like LISA are designed to only segment one object for each question with other objects missed. However, AURA can support multi-object segmentation in a single conversation, responding with textual answers describing various objects with detailed elaborations.

Qualitative Results. Figure 5 presents the visualization results of AURA in three cases, showing that AURA can accurately understand the occlusion relation between objects in the image, reason the purpose underneath the question, and predict segmentation masks for various objects. In the first case, although part of the orange is completely occluded by the ice cream on top, and the plate below has a messy pattern, which may cause confusion, AURA still can correctly understand the question and predict an intact amodal segmentation mask for the occluded orange. In the second case, the black cat and the person's black clothing are easily con-

Index	$\mathcal{E}_{OC}$	$\mathcal{E}_{SO}$	Amodal		Visible	
			gIoU	cIoU	gIoU	cIoU
1			43.55	43.92	48.96	48.63
2	✓		44.86	44.07	49.57	49.10
3		✓	<u>46.95</u>	<u>45.60</u>	<u>50.68</u>	<u>53.72</u>
4	✓	✓	**47.76**	**47.32**	**51.31**	**55.38**

Table 2. Ablation study of designs for occlusion learning. Best in **bold**, second <u>underlined</u>.

Method	Amodal		Visible	
	gIoU	cIoU	gIoU	cIoU
AURA w/o LoRA	20.84	21.03	25.43	27.95
AURA w/ LoRA	**47.76**	**47.32**	**51.31**	**55.38**

Table 3. Ablation study of the effectiveness of using LoRA in the Multi-Modal LLM for efficient fine-tuning. w/o and w/ denote without and with the LoRA technique, respectively. Best in **bold**.

Multi-Modal LLM Backbone	Amodal		Visible	
	gIoU	cIoU	gIoU	cIoU
AURA-7B	47.76	47.32	51.31	55.38
AURA-13B	<u>49.41</u>	<u>49.74</u>	<u>54.13</u>	<u>57.97</u>
AURA-Llama2-13B	**52.34**	**52.95**	**58.42**	**60.21**

Table 4. Ablation study of the scale of the backbone for the Multi-Modal LLM. Best in **bold**, second <u>underlined</u>.

Vision Backbone	Amodal		Visible	
	gIoU	cIoU	gIoU	cIoU
SAM-ViT-B	42.85	42.76	46.92	50.12
SAM-ViT-L	<u>45.23</u>	<u>45.31</u>	<u>49.18</u>	<u>53.71</u>
SAM-ViT-H	**47.76**	**47.32**	**51.31**	**55.38**

Table 5. Ablation study of the scale of the Vision Backbone. B, L, and H denote Base, Large, and Huge, respectively. Best in **bold**, second <u>underlined</u>.

fused, especially in a dark environment. However, AURA can precisely distinguish between the cat and the person's black clothing and accurately predict the complete amodal mask of the person, even though a large region of the person is occluded by the cat. In the third case, the front side of the toilet is occluded by the cat to a large extent, while AURA can reason the purpose of the question and predict a high-quality amodal segmentation mask for the occluded toilet and an accurate segmentation mask for the cat.

5.3. Ablation Study

We conduct ablation studies for the proposed AURA from various aspects, including the design of occlusion learning modules, the effectiveness of LoRA, the scale of the Multi-Modal LLM Backbone, and the scale of the Vision Backbone.

Occlusion Learning. As shown in Table 2, when the two designed Occlusion Condition Encoder $\mathcal{E}_{OC}$ and Spatial Occlusion Encoder $\mathcal{E}_{SO}$ are not used, the performance of AURA drops in a certain degree. When only using $\mathcal{E}_{OC}$, the occlusion conditions of all objects are learned and can help with our model for comprehensively distinguishing whether objects are occluded and the degree of occlusion for these objects. Besides, when only using $\mathcal{E}_{SO}$, the spatial occlusion distribution can be learned and guide our model to be aware of whether each region is occluded or visible, and constrain the consistency between the predicted visible and amodal masks regarding the spatial occlusion condition. Finally, when both $\mathcal{E}_{OC}$ and $\mathcal{E}_{SO}$ are utilized, the best performance is achieved with the learning of global and spatial occlusion conditions simultaneously.

Effectiveness of LoRA. We conduct experiments to evaluate the effectiveness of LoRA, as shown in Table 3. The LoRA [13] technique is utilized to be equipped with the Multi-Modal LLM for efficiently fine-tuning new data while keeping the pre-trained weight of LLM frozen. As shown in Table 3, when not using the LoRA technique, the performance of AURA drops to a certain extent. This is because the Multi-Modal LLM is not pre-trained on the AmodalReasonSeg dataset and has no specific knowledge learned to understand the purpose underneath the user's question and reason for various occluded objects. Therefore, the Multi-Modal LLM cannot predict textual answers correctly to respond to the user's question, leading to the performance of visible and amodal segmentation not well.

Scale of Multi-Modal LLM Backbone. The effect of the scale of the Multi-Modal LLM backbone is evaluated as shown in Table 4. The AURA-7B, AURA-13B, and AURA-Llama2-13B denote the AURA method with different Multi-Modal LLM backbones corresponding to LLaVA-7B, LLaVA-13B, and LLaVA-Llama2-13B models. These models are different variants of the LLaVA model with different scales of parameters (7 billion and 13 billion) and different generations of Llama. All experiments, apart from this ablation study, are based on the AURA-7B model. We can observe from Table 4 that the AURA-7B model does not perform as well as the other two models since the smaller number of parameters limits its learning ability for fine-tuning on the proposed AmodalReasonSeg dataset. Moreover, using larger models with 13 billion parameters can bring better performance since the learning capabilities of these models are better. Finally, upgrading the Llama large language model from the first to the second generation further enhances segmentation performance by generating higher-quality answers and refined embeddings used for segmentation.

Scale of the Vision Backbone. To evaluate the effect of the scale of the Vision Backbone, which is utilized to ex-

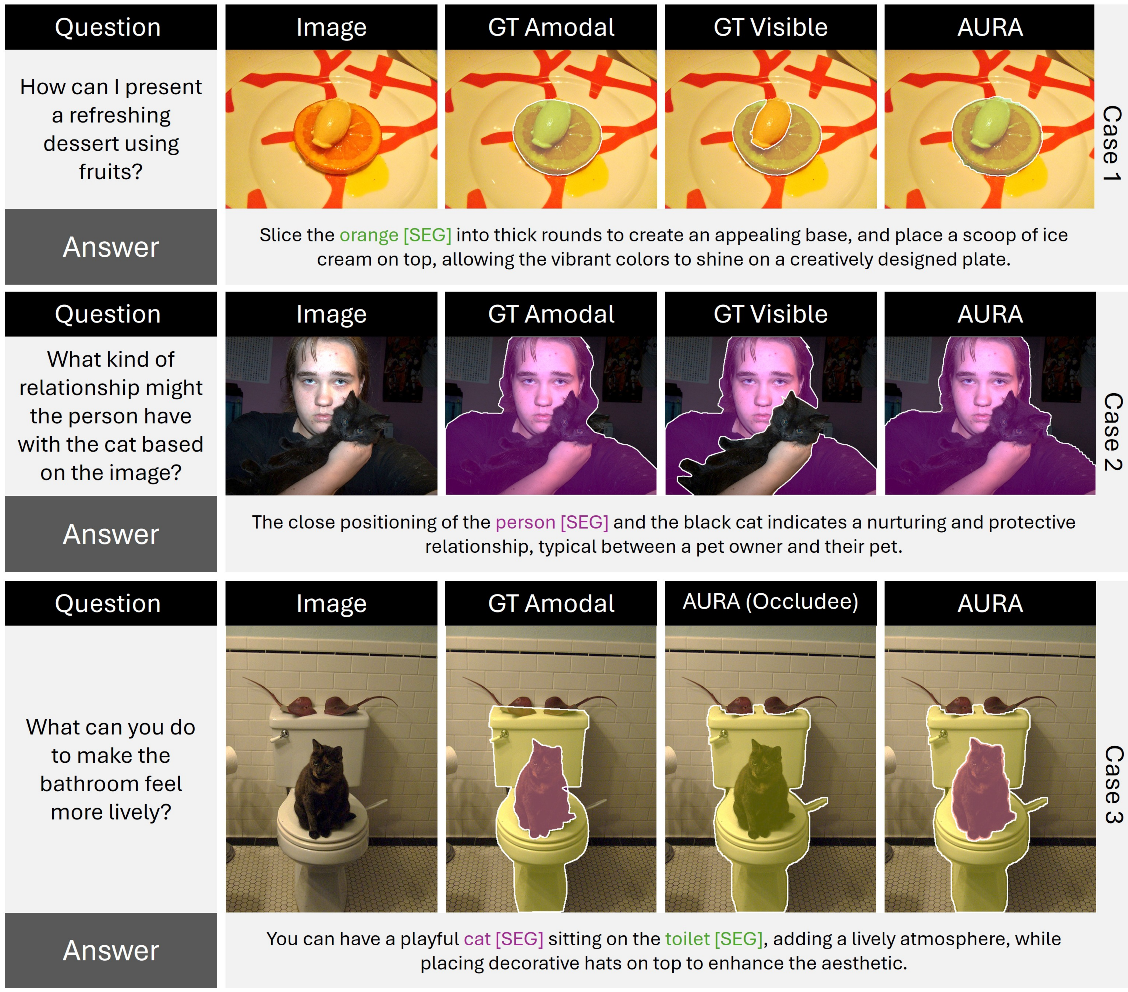

Figure 5. Visualization results of AURA. Three cases are shown from top to bottom with questions and ground-truth (GT) answers. The first two cases show the images, ground-truth amodal and visible masks, and the predicted masks of AURA. In the last case, the predicted amodal mask by AURA for the occludee, which is the occluded toilet, is shown replacing the ground-truth visible mask.

tract vision features of the input image and plays an essential role in providing informative features to the segmentation decoders, experiments are conducted, and the results are shown in Table 5. Three scales of the Vision Backbone are compared, including the Base, Large, and Huge versions of SAM, corresponding to 93.7M, 312M, and 641M parameters, respectively. We can observe that the model with a larger number of parameters can achieve better segmentation performance and greater learning ability.

6. Conclusion

In this work, we propose a new task named Amodal Reasoning Segmentation, aiming to predict textual answers alongside visible and amodal segmentation masks in response to the user's input question that requires understanding and complex reasoning in the image. A generalizable dataset generation pipeline is proposed, which can generate annotations with high quality efficiently. A new dataset named AmodalReasonSeg is presented for daily scenarios where occlusions are challenging for amodal segmentation. Extensive textual question-and-answer pairs are generated, and various objects are referred to in the answers. Moreover, we propose AURA to understand the user's implicit purpose and reason for the answer, which organizes various objects coherently with accurately predicted visible and amodal segmentation masks. Extensive experiments demonstrate the effectiveness of our proposed AURA on reasoning and segmenting the amodal masks in diverse complex scenarios. To enhance the applicability in real-world scenarios, we plan to further optimize the efficiency of our LLM-based model and enrich our dataset's diversity and granularity. Moreover, feature compression approaches [15–17, 40, 41] can be considered to improve the inference speed of the proposed approach.

Acknowledgments

This research was supported by the Ministry of Education of Singapore Tier3 grant MOET32022-0006. This research was also supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Global Talent for Generative AI Copyright Infringement and Copyright Theft, Project Number: RS-2024-00398413, Contribution Rate: 10%).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *International Conference on Robotics and Automation*, pages 5085–5092. IEEE, 2022.

[3] Jasmin Breitenstein, Jonas Löhdefink, and Tim Fingscheidt. Joint prediction of amodal and visible semantic segmentation for automated driving. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 633–645. Springer, 2022.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[5] Chilam Cheang, Haitao Lin, Yanwei Fu, and Xiangyang Xue. Learning 6-dof object poses to grasp category-level objects by language instructions. In *International Conference on Robotics and Automation*, pages 8476–8482. IEEE, 2022.

[6] Junjie Chen, Li Niu, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Amodal instance segmentation via prior-guided expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 313–321, 2023.

[7] Helisa Dhamo, Nassir Navab, and Federico Tombari. Object-driven multi-layer scene decomposition from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5369–5378, 2019.

[8] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6144–6153, 2018.

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[10] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1328–1336. IEEE, 2019.

[11] Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Coarse-to-fine amodal segmentation with shape prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1262–1271, 2023.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2021.

[14] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. SAIL-VOS: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3105–3115, 2019.

[15] Chen Hui, Shengping Zhang, Wenxue Cui, Shaohui Liu, Feng Jiang, and Debin Zhao. Rate-adaptive neural network for image compressive sensing. *IEEE Transactions on Multimedia*, 26:2515–2530, 2023.

[16] Chen Hui, Haiqi Zhu, Shuya Yan, Shaohui Liu, Feng Jiang, and Debin Zhao. S^2-csnet: Scale-aware scalable sampling network for image compressive sensing. In *ACM Multimedia*, pages 1–10, 2024.

[17] Chen Hui, Debin Zhao, Weisi Lin, Shaohui Liu, and Feng Jiang. Image compressive sensing with scale-variable adaptive sampling and hybrid-attention transformer reconstruction. *IEEE Transactions on Multimedia*, 2025.

[18] Xueying Jiang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Domain generalization via balancing training difficulty and model capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18993–19003, 2023.

[19] Xueying Jiang, Sheng Jin, Lewei Lu, Xiaoqin Zhang, and Shijian Lu. Weakly supervised monocular 3D detection with a single-view image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10508–10518, 2024.

[20] Xueying Jiang, Sheng Jin, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Monomae: Enhancing monocular 3D detection through depth-aware masked autoencoders. In *arXiv preprint arXiv:2405.07696*, 2024.

[21] Xueying Jiang, Lewei Lu, Ling Shao, and Shijian Lu. Multimodal 3d reasoning segmentation with complex scenes. *arXiv preprint arXiv:2411.13927*, 2024.

[22] Xueying Jiang, Wenhao Li, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Exploring 3d activity reasoning and planning: From implicit human intentions to route-aware planning. *arXiv preprint arXiv:2503.12974*, 2025.

[23] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary ob-

ject detection with fine-grained descriptors. In *International Conference on Learning Representations*, 2024.

[24] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *Advances in Neural Information Processing Systems*, 2024.

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

[26] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

[28] Ke Li and Jitendra Malik. Amodal instance segmentation. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 677–693. Springer, 2016.

[29] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. 2D amodal instance segmentation guided by 3D shape prior. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 165–181. Springer, 2022.

[30] Zhixuan Li, Ruohua Shi, Tiejun Huang, and Tingting Jiang. OAFormer: Learning occlusion distinguishable feature for amodal instance segmentation. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1–5. IEEE, 2023.

[31] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. GIN: Generative invariant shape prior for amodal instance segmentation. *IEEE Transactions on Multimedia*, 2023.

[32] Zhixuan Li, Weining Ye, Juan Terven, Zachary Bennett, Ying Zheng, Tingting Jiang, and Tiejun Huang. MUVA: A new large-scale benchmark for multi-view amodal instance segmentation in the shopping scenario. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23504–23513, 2023.

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

[34] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. *Advances in Neural Information Processing Systems*, 33: 16246–16257, 2020.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Zhaochen Liu, Zhixuan Li, and Tingting Jiang. BLADE: Box-level supervised amodal segmentation through directed expansion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[38] Zhaochen Liu, Limeng Qiao, Xiangxiang Chu, and Tingting Jiang. PLUG: Revisiting amodal segmentation with foundation model and hierarchical focus. *arXiv preprint arXiv:2405.16094*, 2024.

[39] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] Yachun Mi, Yan Shu, Yu Li, Chen Hui, Puchao Zhou, and Shaohui Liu. Clif-vqa: Enhancing video quality assessment by incorporating high-level semantic information related to human feelings. In *Proceedings of the ACM International Conference on Multimedia*, pages 9989–9998, 2024.

[41] Yachun Mi, Yu Li, Weicheng Meng, Chaofeng Chen, Chen Hui, and Shaohui Liu. Mvqa: Mamba with unified sampling for efficient video quality assessment. *arXiv preprint arXiv:2504.16003*, 2025.

[42] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21023–21032, 2022.

[43] Khoi Nguyen and Sinisa Todorovic. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7396–7405, 2021.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[45] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with KINS dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019.

[46] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[47] N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9356–9366, 2022.

[48] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[49] Ahmed Rida Sekkat, Rohit Mohan, Oliver Sawade, Elmar Matthes, and Abhinav Valada. Amodalsynthdrive: A synthetic amodal perception dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 9(11):9597–9604, 2024.

[50] Mennatullah Siam, Sara Elkerdawy, Martin Jagersand, and Senthil Yogamani. Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *IEEE international conference on intelligent transportation systems*, pages 1–8. IEEE, 2017.

[51] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[52] Yihong Sun, Adam Kortylewski, and Alan Yuille. Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2022.

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[55] Junchi Wang and Lei Ke. LLM-Seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1765–1774, 2024.

[56] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024.

[57] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. Amodal segmentation based on visible region segmentation and shape prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2995–3003, 2021.

[58] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024.

[59] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9109, 2024.

[60] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 2024.

[61] Ma Yi-de, Liu Qing, and Qian Zhi-Bai. Automated image segmentation using improved pcnn model based on cross-entropy. In *Proceedings of International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 743–746. IEEE, 2004.

[62] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[63] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3792, 2020.

[64] Jingyi Zhang, Jiaxing Huang, Xueying Jiang, and Shijian Lu. Black-box unsupervised domain adaptation with bi-directional atkinson-shiffrin memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11771–11782, 2023.

[65] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *International Conference on Learning Representations*, 2024.

[66] Shihui Zhang, Ziteng Xue, Yuhong Jiang, and Houlin Wang. Opnet: Deep occlusion perception network with boundary awareness for amodal instance segmentation. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2595–2599. IEEE, 2024.

[67] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37:71737–71767, 2024.

[68] Ziheng Zhang, Anpei Chen, Ling Xie, Jingyi Yu, and Shenghua Gao. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In *ACM International Conference on Multimedia*, pages 2124–2132, 2019.

[69] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[70] Chuanxia Zheng, Duy-Son Dao, Guoxian Song, Tat-Jen Cham, and Jianfei Cai. Visiting the invisible: Layer-by-layer completed scene decomposition. *International Journal of Computer Vision*, 129:3195–3215, 2021.

[71] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017.