Supplementary Material of 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

In the supplementary material, we introduce the following contents including:

- Generalization ability is demonstrated in Section 1,
- Experimental results in Section 2,
- Discussions in Section 3,
- Additional Information of the Dataset in Section 4.

1. Generalization on Cloud Occluded Region Segmentation

We showcase the proposed method's potential of generalizing to the meteorological image, which is included to qualitatively demonstrate the potential of our method to generalize across domains. The example is illustrative with no quantitative results, and is not part of the core evaluation presented in the main paper.

To be specific, cloud cover is a major challenge in remote sensing and meteorological observations, as it obscures ground-level features that are critical for weather modeling, environmental monitoring, and climate studies. Accurate retrieval of ground information under cloud cover is essential for enhancing the quality of meteorological data and improving predictions.

Ground-level observations, such as vegetation health, urban heat distribution, and soil moisture, play a vital role in meteorological models. However, cloud coverage often introduces uncertainties in these observations, affecting the accuracy of derived climate indicators. Besides, meteorological models rely on ground information for initial boundary conditions, such as vegetation cover for evapotranspiration modeling or urban features for heat island studies. Addressing cloud-induced occlusions can significantly improve the reliability of such models. Furthermore, during extreme weather events like tropical cyclones or heavy rainfall, cloud cover is prevalent. Recovering occluded ground-level features aids in post-disaster assessment and supports real-time meteorological decision-making.

*Corresponding author.

It is worth noticing that the dynamic nature of cloud cover, influenced by meteorological parameters such as wind speed, humidity, and atmospheric pressure, makes it critical to develop robust methods that can adaptively infer ground information under varying cloud conditions.

We evaluate the proposed AURA method for reconstructing ground-level information obscured by cloud cover using amodal segmentation, enabling meteorologists to retrieve critical land-surface data under cloudy conditions.

As shown in Figure 1, the effectiveness of AURA is evaluated using a cloud-occluded image from the RICE dataset [1]. This dataset is derived from the Landsat 8 OLI/TIRS dataset [2], which comprises paired images captured under cloud-occluded and non-occluded conditions. The Landsat 8 OLI/TIRS dataset, developed by NASA and the United States Geological Survey (USGS), offers high-resolution multispectral and thermal imagery to support global environmental monitoring, climate change research, and sustainable resource management. For this study, we annotated ten images from the RICE dataset with ground-truth amodal masks and question-answer pairs. AURA was fine-tuned on eight images, reserving the remaining two for validation. The same model structure and parameter settings of AURA are used, as described in Section 5 of the main paper.

Figure 1 presents a qualitative result on the validation set. The results highlight AURA's capability to segment complete shapes despite cloud occlusions, demonstrating its potential to comprehend ground-level conditions based on meteorological knowledge.

2. Experiments

2.1. Experimental Settings

All of the experiments are conducted on the proposed AmodalReasonSeg dataset. No extra datasets other than the proposed AmodalReasonSeg dataset are used to ensure fairness in comparison. There are no other datasets used in our experiments. The pre-trained weights of LLaVA-7B-v1-1

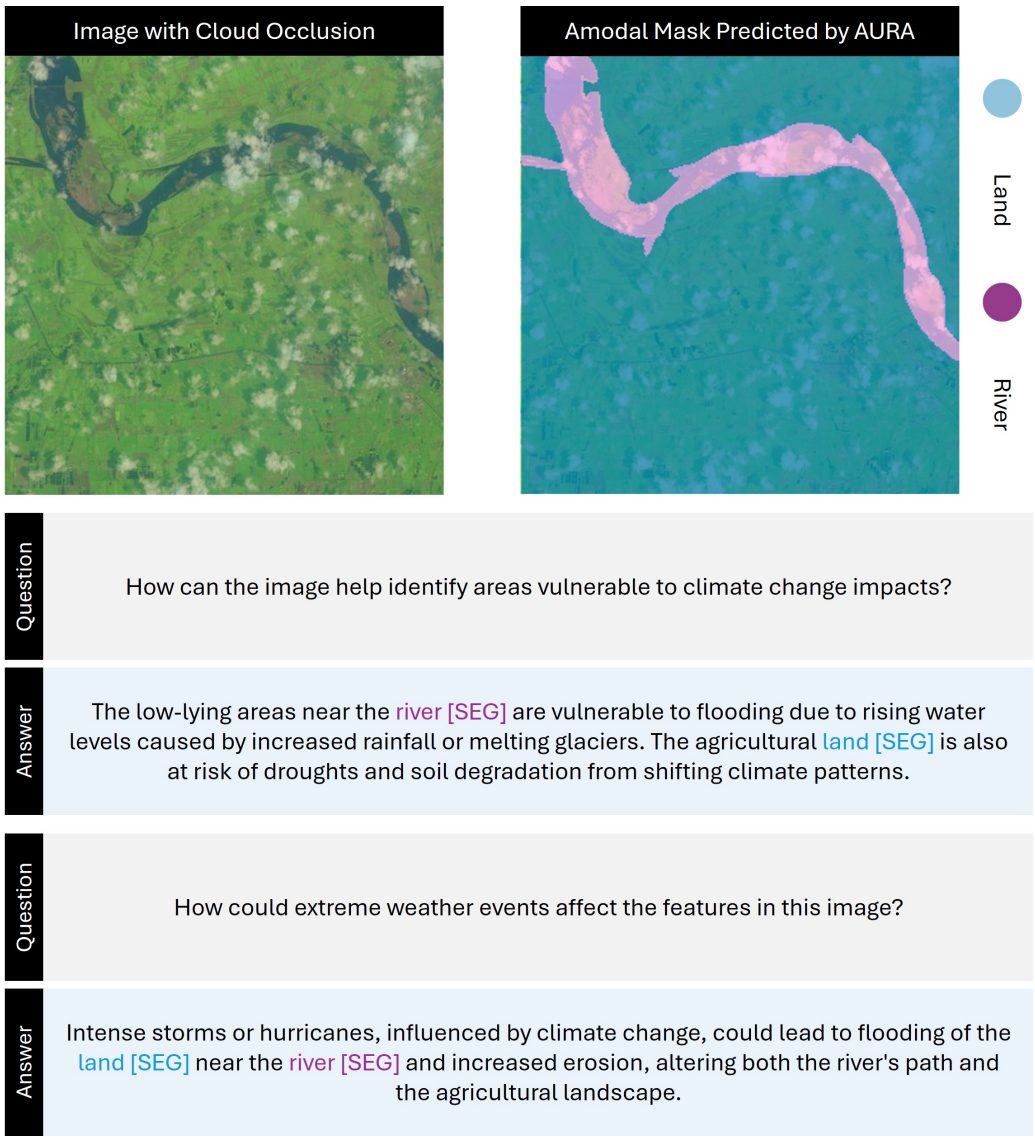


Figure 1. The qualitative result for demonstrating the generalization ability of AURA on real-world scenarios such as climate change analysis. Best viewed in color.

and SAM-Vit-H are used as the initialized weights of the multi-modal LLM as well as the vision backbone and mask decoders, respectively.

2.2. Training and Inference Efficiency

Regarding training time, LISA requires 8 hours for either visible or amodal segmentation, whereas AURA takes only 10 hours to handle both.

Under the same testing environment, LISA achieves 0.33 fps, whereas AURA achieves 0.28 fps.

2.3. Evaluation Details

Evaluation of amodal segmentation methods. The evaluation of these models is performed by computing metrics between all predicted objects and the target object, as these models are designed to predict all objects in the scene and cannot identify which specific object should be predicted.

Explanation of the evaluation of MLLM-based methods. First, MLLM methods like LISA and LLM-Seg predict one object at a time and are evaluated per prediction–target match, ignoring any unmatched targets. Because most images contain multiple objects, these methods tend to perform suboptimally. Second, for MLLM-based methods that can predict multiple objects simultaneously, such as GVSA, OMG-LLaVA, GLaMM, and PSALM, the evaluation protocol is consistent with that used for AURA.

3. Discussions

3.1. Difference between Reasoning Segmentation and Referring Segmentation

The key difference between the reasoning segmentation and the referring segmentation tasks lies in the content of the questions and answers.

In the reasoning segmentation task, questions are abstract and do not explicitly reference specific objects. The answers include both textual elaborations and segmentation masks for the objects mentioned. For example, the question, "How can I accessorize my kitchen for cooking efficiently?" requires understanding and reasoning without direct object references. The corresponding answer might be, "Utilize the bottles[SEG] on the countertop for easy access to oils or spices, and rely on the sink[SEG] for rinsing utensils and ingredients during food preparation." This answer provides both the segmentation masks (denoted as "[SEG]") and explanatory details to address the question.

In contrast, the referring segmentation task involves questions that explicitly mention the objects to be segmented. For instance, the question, "Where is the orange? Please segment it.", explicitly identifies the object of interest. The typical answer is concise, such as "It is [SEG]", and lacks additional elaborations.

3.2. Difference between Amodal Reasoning Segmentation and other Amodal Segmentation Tasks

The proposed amodal reasoning segmentation task differs from existing amodal segmentation tasks by enabling interaction with users through textual input questions and providing elaborative textual answers.

While traditional amodal segmentation tasks—spanning semantic, instance, and panoptic levels—are limited to processing only image or video inputs, the proposed task introduces the ability to accept textual questions as input, infer the implicit purpose underneath these questions, and produce textual answers alongside segmentation masks. The ability of language interaction significantly broadens the potential for real-world applications.

3.3. Applicability of Existing Reasoning Segmentation Method on the Proposed New Task

There are various existing reasoning segmentation methods proposed, like LISA. These methods were originally designed for segmenting visible masks, and can be applied to the proposed new task - Amodal Reasoning Segmentation by training with ground-truth question-answer pairs and amodal segmentation masks.

However, methods like LISA need two major improvements for this task. First, LISA can only predict visible or amodal masks, requiring an additional decoder to handle both simultaneously. Second, LISA predicts only one object, while AURA handles multiple objects in a single inference by modifying [SEG] tokens. With these changes, LISA can only be upgraded to the "Baseline" method in Table 1 of our main paper.

Then the task-specific designs, including the proposed Occlusion Condition Encoder and the Spatial Occlusion Encoder, are required to handle the occlusion problem, which is the core difference between the traditional visible segmentation task and the focused amodal segmentation task in the paper.

[Core Task Description] Your responsibility is to generate questions and respective answers based on the content of a provided image. The dimensions of the image are {image_height} pixels in height and {image_width} pixels in width. Your goal is to craft ten question-and-answer pairs for each image, designed to push a Large Vision-Language Model to perform a detailed analysis and interpretation of the visual content.

[Guidelines for Question and Answer Creation] Please adhere to the following principles while formulating questions and answers:
1.The questions should be brief, specific, and closely tied to the image.
2.Answers must draw exclusively from the objects listed in the provided (Image Details) section.
3.Avoid referencing the name or location of any object within the question itself.
4.Frame questions around an object's attributes or functionality.
5.Ensure the questions directly relate to the image's context, avoiding generic inquiries.
6.Limit repetition: no more than two different questions should lead to the same answer, and each question should offer a unique perspective.
7.When possible, ask about complete activities depicted in the image.
8.Aim to incorporate as many objects from the provided list into the answers as possible.
9.Provide reasoning within the answers. For answers involving multiple objects, give unique explanations for each. Include the category name or object name for each object in the answer, but avoid mentioning them in the question.
10.Where applicable, integrate information about occlusions between objects into the answers.
11.Format all question-and-answer pairs as follows: "1. Question: ... Answer: ...," replacing "1" with the corresponding sequence number.

Here's an example:
•**Question:** "What is a good way to spend a relaxing afternoon at home?"
•**Answer:** "You can recline on the comfortable couch, place your glasses on, grab the book lying open on the side table, sip a warm cup of coffee, and enjoy the company of your dog curled up beside you."

[Image Details] Additional data about the image is provided below. The image includes {object_count} object(s) with detailed information outlined as follows:
Each object is represented in the format:
•**Object ID:** <object_ID>
•**Category ID:** <category_ID>
•**Category Name:** <category_name>
•**Bounding Box:** <bounding_box>
The bounding box specifies the coordinates of the amodal mask (both occluded and visible portions) for the object, presented as [top-left x, top-left y, bottom-right x, bottom-right y]. Below is the detailed list of objects in the image:
[Provide object details here.]

Figure 2. The designed prompt template used for guiding the GPT, including the core task description for the ChatGPT-4o, the guidelines for question and answer creation, and the details about the image and objects in it.

4. Dataset

4.1. Exhibition of Various Perspectives

In the created dataset AmodalReasonSeg, there are 11.3 pairs of questions and answers annotated for each image on average. To ensure the diversity of the question-and-answer pairs provided in this dataset, we specifically add a requirement in the prompt template used for guiding ChatGPT-4o to generate question-and-answer pairs from different per-

spectives to cover various potentials. As shown in Figures 3 and 4, two cases of the proposed AmodalReasonSeg dataset are shown to exhibit the diversity of question-and-answer pairs in this dataset.

4.2. Prompt Template for GPT

We present the designed prompt template in Figure 2. This template specifies detailed requirements and provides comprehensive information about the image and objects, ensuring the generation of high-quality questions and answers by ChatGPT-4o. Additionally, it includes an example to clarify the requirements and assist ChatGPT-4o in understanding them effectively.

4.3. Average count of question-answer pairs per image.

The dataset contains an average of 11.3 QA pairs per image. Initially, GPT-4o generated 10 QA pairs per image, which human annotators subsequently revised and occasionally supplemented during verification to enhance data quality, resulting in a final average above 10.

References

[1] Daoyu Lin, Guangluan Xu, Xiaoke Wang, Yang Wang, Xian Sun, and Kun Fu. A remote sensing image dataset for cloud removal. In *arXiv preprint arXiv:1901.00600*, 2019. 1

[2] David P Roy, Michael A Wulder, Thomas R Loveland, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Dennis Helder, James R Irons, David M Johnson, Robert Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of Environment*, 145: 154–172, 2014. 1

Image

Ground-truth Amodal Mask

Question: What items are essential for maintaining hygiene in a bathroom setting?

Answer: You can use the sink [SEG] to wash your hands and face, and the toilet [SEG] for personal needs, ensuring cleanliness and hygiene in the space.

Question: How can I create a relaxing atmosphere while preparing for a bath?

Answer: You can close the shower curtain for privacy, use the sink [SEG] to set out toiletries [SEG], and have towels ready for drying off after your bath.

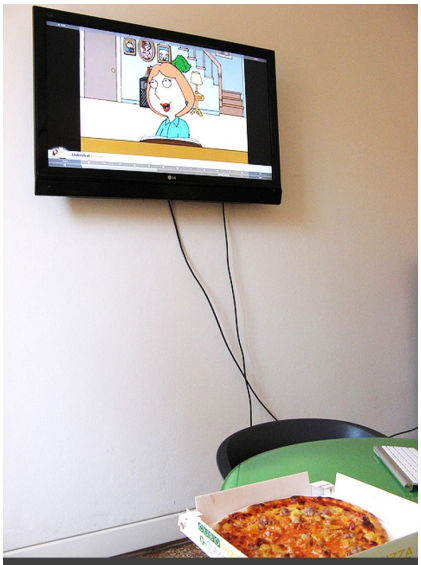
Question: What features can I include in a compact bathroom for efficient use of space?

Answer: Using a compact toilet [SEG], a small sink [SEG], and a shower curtain saves space while providing essential functions without overcrowding the area.

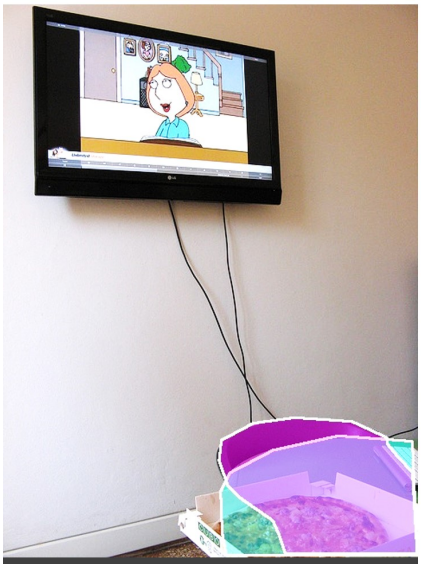
Question: How can I effectively store items in a limited bathroom space?

Answer: Utilize the counter by placing toiletries [SEG] near the sink [SEG] for easy access, and keep a towel rack nearby for quick drying after uses, ensuring everything is within reach.

Figure 3. The first example of the proposed AmodalReasonSeg dataset shows various questions and answers from diverse perspectives for the same image. Best viewed in color.



Image



Ground-truth Amodal Mask

Question	How can you make your dining experience more enjoyable while engaged with entertainment?
Answer	You can sit comfortably at the chair [SEG] next to the dining table [SEG], grab a slice of pizza from the box below, and watch the amusing scenes playing on the TV, making the combination perfect for leisure.
Question	What combination can create a cozy atmosphere for a casual gathering?
Answer	The TV showing a lively cartoon can provide entertainment, while the pizza box on the dining table [SEG] offers tasty snacks, and sitting in the chair [SEG] can make the gathering feel relaxed and enjoyable.
Question	How can you make the most of a relaxed evening at home?
Answer	You can sit in the chair [SEG] and share pizzas from the box on the dining table [SEG] while enjoying a captivating show on the TV, creating a perfect setting for relaxation and fun.
Question	What arrangement supports a relaxing night filled with good food and fun?
Answer	Sitting at the chair [SEG] by the dining table [SEG] with a box of pizza while watching a fun animated series on the TV creates a perfectly cozy setup.

Figure 4. The second example of the proposed AmodalReasonSeg dataset shows various questions and answers from diverse perspectives for the same image. Best viewed in color.