1 Implementation Details

Image Occlusion. For each high-causality entity, we first utilize GLIP [Li *et al.*, 2022] to ground the corresponding object in the image, generating a bounding box. We then input this bounding box into SAM [Kirillov *et al.*, 2023] to obtain a mask that precisely delineates the object's region. To seamlessly apply our method across different LVLM architectures, we directly apply occlusion to the segmented pixels. Specifically, we utilize square patches to occlude the object, randomly placing these patches within the object's mask with a small interval. To enhance the occlusion effect on objects of varying sizes, the patch size is set to one-third of the shortest side of the object's bounding box. Additionally, the patch is filled with the mean color of ImageNet [Deng *et al.*, 2009]. We illustrate several examples in Figure 1. The complete implementation is provided in our code repository.

Answer Checking. In trial and error learning, we need to determine whether the extracted answer aligns with the target answer. To accomplish this, we assess answer correctness through semantic similarity using BGE-M3 [Chen *et al.*, 2024]. The implementation details are available in the provided code.

2 Prompting Examples

Our method relies on several prompting examples to elicit the generation from language models through in-context learning. Here, we provide detailed prompting examples used for extracting entities from captions (Table 1), generating specialized instructions for entities (Table 2), and extracting answers from sampled rationales (Table 3).

3 Benchmark Leaderboards

We obtain the results for GPT-4V [Achiam *et al.*, 2023] and Gemini Pro [Team *et al.*, 2023] on 4 comprehensive benchmarks from their respective official leaderboards. The URL for the leaderboard of each comprehensive benchmark are listed in Table 4.

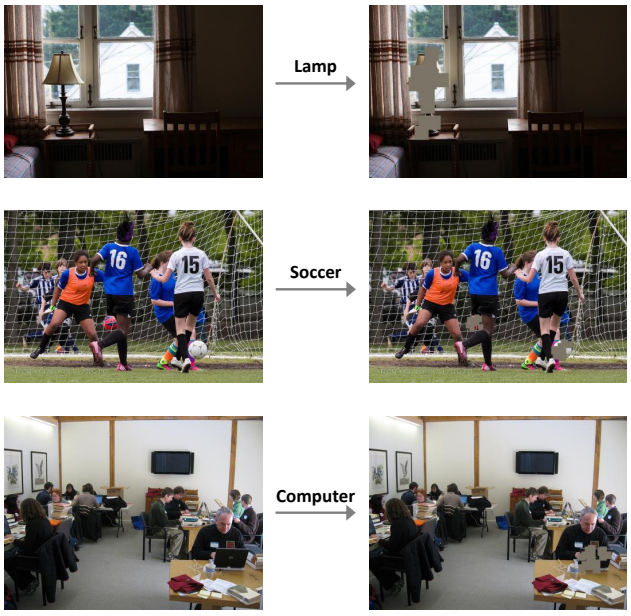

Figure 1: Several examples of occluded images.

References

[Achiam *et al.*, 2023] 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.

[Chen *et al.*, 2024] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[Kirillov *et al.*, 2023] 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.

[Li *et al.*, 2022] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[Li *et al.*, 2023] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[Liu *et al.*, 2023] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Yu *et al.*, 2023] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

```
You are an entity extractor that helps me extract entities from text.  I will provide you
with a piece of text, which is a description of an image.  Since I need to get the textual
descriptions of the entities that appear in the image, I need you to extract entities from
the given text as accurately as possible.

Here are some definitions of entity:
1) Entities are concrete objects that can be recognized in the corresponding image (such as
objects, people, places, products).
2) The entities to be extracted do not need to be named entities; as long as they are
concrete objects, they can be extracted.
3) Entities can appear in text in various forms, including single word, noun phrases.
4) For abstract concept nouns or noun phrases (such as emotions, states, qualities,
sensations, actions), they are not considered as entities.

Here are some constraints you need to follow:
1) The form of entities does not consider any format other than single word and noun
phrases.
2) If a noun phrase is not an entity, then the nouns within that phrase cannot be considered
as entities either.
3) To extract an entity, you first need to identify the complete expression and then remove
any modifiers.  The text format is as follows:  "1.  entity w/ modifiers -> entity w/o
modifiers".
4) Please do not extract modifiers indicating quantity.

Here are some examples:

Text:  People sunbathing and sitting under umbrellas at a city beach.  Extracted entities:
<begin>
1.  people -> people
2.  umbrellas -> umbrellas
3.  city beach -> city beach
<end>

Text:  The image is a collage of various kitchen items, including a clock, pots, pans, a
refrigerator, and a chalkboard.
Extracted entities:
<begin>
1.  clock -> clock
2.  pots -> pots
3.  pans -> pans
4.  refrigerator -> refrigerator
5.  chalkboard -> chalkboard
<end>

Text:  A male tennis player hits the ball on a grass court.  Extracted entities:
<begin>
1.  male tennis player -> tennis player
2.  ball -> ball
3.  grass court -> grass court
<end>

Text:  A group of people walks down a train platform, with a yellow train stopped nearby.
Extracted entities:
<begin>
1.  people -> people
2.  train platform -> train platform
3.  yellow train -> train
<end>

...

Text:  A motorbike parked on a roadside close to some bush.
Extracted entities:
<begin>
1.  motorbike -> motorbike
2.  bush -> bush
<end>
```

Table 1: Detailed prompting examples for entity extraction.

```
You are a question constructor.  I will provide you with an entity, which could refer to a
specific object, a specific type of object, a place, a location or an occupation.
You first need to determine the type of the entity and generate a question according to the
type of the given entity.

Here are some constraints you need to follow:
1) If an object is very common, you can simply refer to it as "object".
2) If an object is relatively uncommon, you may need to specify its type in the question.
3) If an object's type is difficult to determine, you can simply refer to it as "object".
4) Should not reveal the entire entity's information in the generated question.

Here are some examples:

Entity:  sewing machine
<begin>
Question:  In the given image, there is a machine that is heavily occluded by a cluster of
gray blocks.  Please answer the following question.
What kind of machine might the object occluded by a cluster of gray blocks be?  Please
provide your reasoning process and confirm a unique answer.
<end>

Entity:  refrigerator
<begin>
Question:  In the given image, there is an appliance that is heavily occluded by a cluster
of gray blocks.  Please answer the following question.
What might the appliance occluded by the gray blocks be?  Please provide your reasoning
process and confirm a unique answer.
<end>

Entity:  fireman
<begin>
Question:  In the given image, there is a person engaged in a certain profession, who is
heavily occluded by a cluster of gray blocks.  Please answer the following question.
What profession might this partially occluded person be engaged in?  Please provide your
reasoning process and confirm a unique answer.
<end>

Entity:  farm
<begin>
Question:  In the given image, there is a group of person occluded by a cluster of gray
blocks.  Please answer the following question.
What activity might the people occluded by the gray blocks be doing?  Please provide your
reasoning process and confirm a unique answer.
<end>

Entity:  rider
<begin>
Question:  In the given image, there is a view of a place which is heavily occluded by a
cluster of gray blocks.  Please answer the following question.
The view obscured by the gray blocks could be which place?  Please provide your reasoning
process and confirm a unique answer.
<end>

...

Entity:  couple
<begin>
Question:  In the given image, there is a group of person occluded by a cluster of gray
blocks.  Please answer the following question.
What relationship might the people occluded by the gray blocks have?  Please provide your
reasoning process and confirm a unique answer.
<end>
```

Table 2: Detailed prompting examples for instruction generation.

```
You are an answer extractor.  I will provide you with a text, which is an analysis of
reasoning about what an occluded object is.  Please extract the confirmed answers from this
reasoning process.

To extract the answer, here are some constraints you need to follow:
1) If the reasoning process does not yield an answer, please determine it as "unknown".
2) If the reasoning process expresses uncertainty or cannot be determined, please determine
it as "unknown".
3) Please directly provide the extracted answer, and do not include your analysis reasons.

Here are some examples:

Text:  Given the heavily occluded object in the image, it's likely to be a piece of food,
possibly a meat, which is covered in a generous amount of toppings, such as onions and
cheese.  However, without seeing more details, the object's precise type and presence of
other ingredients cannot be confirmed.
<begin>
Extracted Answer:  meat
<end>

Text:  While looking at the image, I noticed that there's a boy in a red beanie and red
jacket sitting in the back seat of the bus.  The partially visible object that is heavily
blocked by gray blocks might be a window.  The initial assumption might be that the boy is
looking out of the window, which could be obstructed by the gray blocks.  However, since
the image also shows another person looking out the window, it's more likely that the gray
blocks are there to provide privacy or to separate sections of the bus.
<begin>
Extracted Answer:  window
<end>

Text:  Based on the visual information available in the image, it is impossible to determine
the exact nature of the object that is heavily occluded by the gray blocks.  However, we
can make some educated guesses.  One possibility is that the object is a monitor or computer
screen, as there is already a computer monitor present in the image.  Another possibility is
that the object is a lamp or other light source, as there is a lamp present in the image.
The gray blocks could also be other random objects, but without more context or visual
clues, it is not possible to identify the object with certainty.
<begin>
Extracted Answer:  monitor, computer screen, lamp
<end>

Text:  Based on the image, the heavily occluded object is in a living room area surrounded
by brown furniture and plants.  Since there is a couch and a chair present in the scene,
it's possible that the object is part of the seating arrangement, such as a bench or an
ottoman.  The gray blocks could be used as an additional seating or as a side table, which
could explain the obscured view of the object.  The reasoning process involves identifying
the living room elements, such as the couch and the chair, and considering the possibility
of an additional seating or side table made up of gray blocks.  <begin>
Extracted Answer:  bench, ottoman
<end>

Text:  The object heavily obscured by the gray blocks might be a white refrigerator or
a microwave placed under a cabinet.  The gray blocks are blocking any clear view of the
object, making it difficult to determine its exact nature.  However, these are two common
appliances found in kitchens, and their size and color match the description of the heavily
blocked object in the image.
<begin>
Extracted Answer:  refrigerator, microwave
<end>

...

Text:  The heavily-occluded object in the image is a flat-screen TV mounted on the wall.
<begin>
Extracted Answer:  flat-screen TV
<end>
```

Table 3: Detailed prompting examples for answer extraction.

Benchmark	Leaderboard URL
MME [Fu *et al.*, 2023]	https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation
MMBench [Liu *et al.*, 2023]	https://mmbench.opencompass.org.cn/leaderboard
SEEDBench [Li *et al.*, 2023]	https://huggingface.co/spaces/AILab-CVC/SEED-Bench_Leaderboard
MM-Vet [Yu *et al.*, 2023]	https://paperswithcode.com/sota/visual-question-answering-on-mm-vet

Table 4: The leaderboard URL of each comprehensive benchmark.