OW-VISCapTor: Abstractors for Open-World Video Instance Segmentation and Captioning

Anwesa Choudhuri, Girish Chowdhary, Alexander Schwing
University of Illinois at Urbana-Champaign

Wed, Dec 11, Poster Session 2 (4:30 - 7:30 p.m. PST)

Overview

New task:
Open-World Video Instance Segmentation and Captioning
(OW-VISCap)

New task:
Open-World Video Instance Segmentation and Captioning
(OW-VISCap)

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

New task:
Open-World Video Instance Segmentation and Captioning
(OW-VISCap)

Detect, segment and track objects across frames

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

New task:
Open-World Video Instance Segmentation and Captioning
(OW-VISCap)

Detect, segment and track objects across frames

Describe the objects with **rich captions**

New task:
Open-World Video Instance Segmentation and Captioning (OW-VISCap)

Detect, segment and track objects across frames

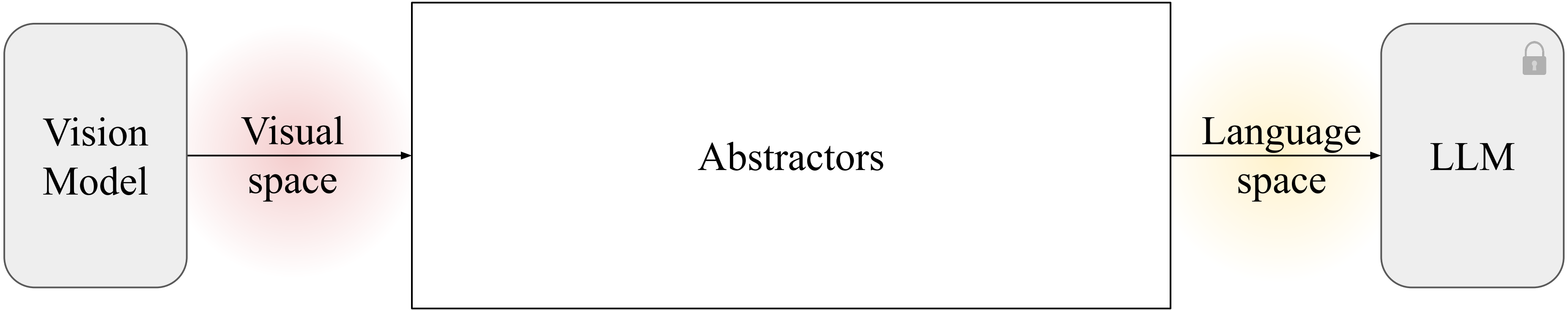
Describe the objects with **rich captions**

For both seen (closed-world) or never before seen (open-world) objects

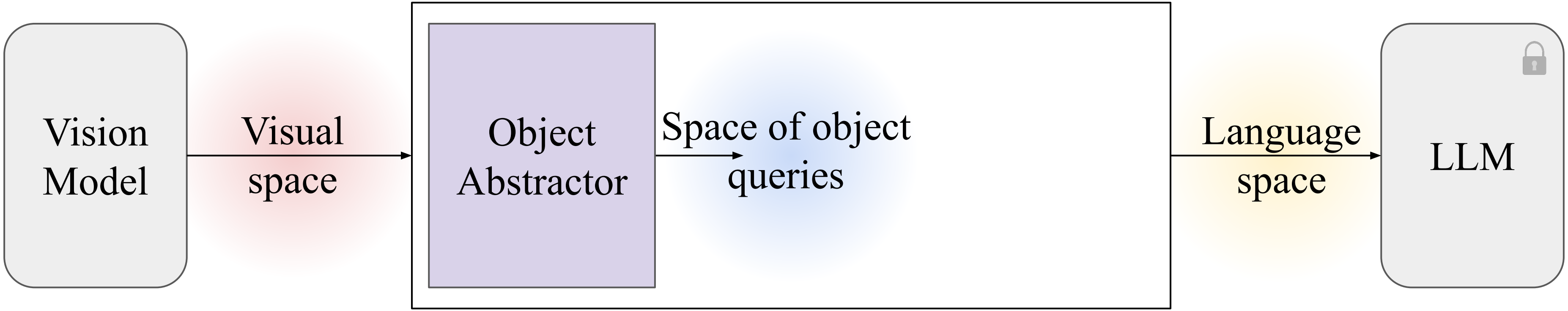
OW-VISCap: Addressed by Abstractors

Abstractors

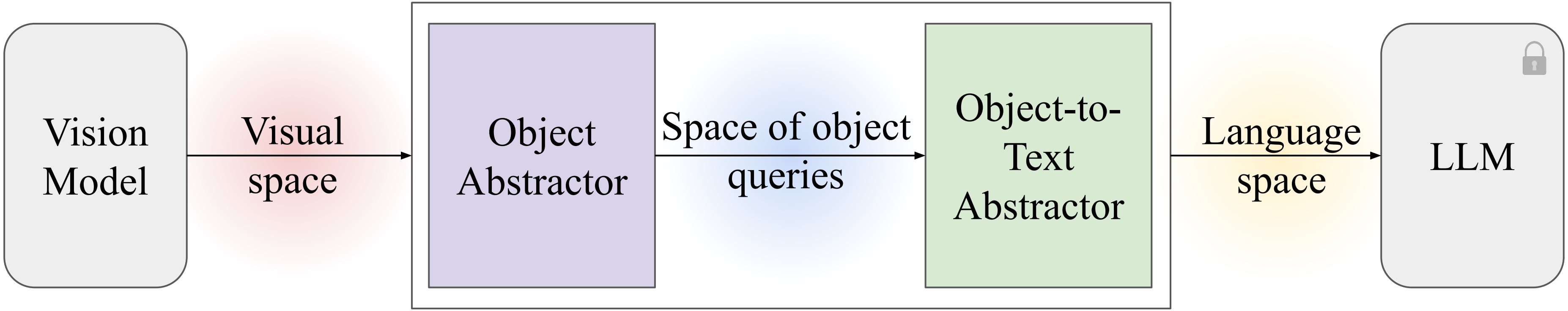
OW-VISCap: Addressed by Abstractors



Addressed by Developing Abstractors



Addressed by Developing Abstractors



OW-VISCapTor: Evaluation

Open-World Video Instance Segmentation
OWTA improved by 5.6 points

Dense Video Object Captioning
CapA improved by 7.1 points

Motivation

[1] Choudhuri et al., CVPR 2023
[2] Huang et al., NeurIPS 2022
[3] Wang et al., CVPR 2021

Prior Work on VIS

Assigns one word label to segmented objects [1, 2, 3] in the closed world

car, car, pedestrian, pedestrian, pedestrian, pedestrian, pedestrian, pedestrian, pedestrian

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Prior Work on VIS

[1] Choudhuri et al., CVPR 2023
[2] Huang et al., NeurIPS 2022
[3] Wang et al., CVPR 2021

Assigns one word label to segmented objects [1, 2, 3] in the closed world

One word labels convey a limited information

car, car, pedestrian, pedestrian, pedestrian, pedestrian, pedestrian, pedestrian, pedestrian

Prior Work on Captioning

[1] Jin et al., NeurIPS 2022
[2] Li et al., arXiv 2023

Video-level or image-level captioning [1, 2]

A street with people walking and cars driving

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

[1] Jin et al., NeurIPS 2022
[2] Li et al., arXiv 2023

Prior Work on Captioning

Video-level or image-level captioning [1, 2]

Doesn't capture object-centric details

A street with people walking and cars driving

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

New Task: OW-VISCap

a car is driving down the street
a car driving down the street
a woman walking down the street
a man with a crutch crossing the street
a woman is standing at a red table by the side of a street
a trash can by the side of a street

...

New Task: OW-VISCap

a man with a crutch crossing the street

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

OW-VISCapTor to Address OW-VISCap

AbstracTors for Open-World Video Instance Segmentation and Captioning

OW-VISCapTor to Address OW-VISCap

AbstracTors for Open-World Video Instance Segmentation and Captioning

Networks that project information from one space to another

Abstractors for OW-VISCap: Challenges

- Haven’t been explored to connect object and language spaces

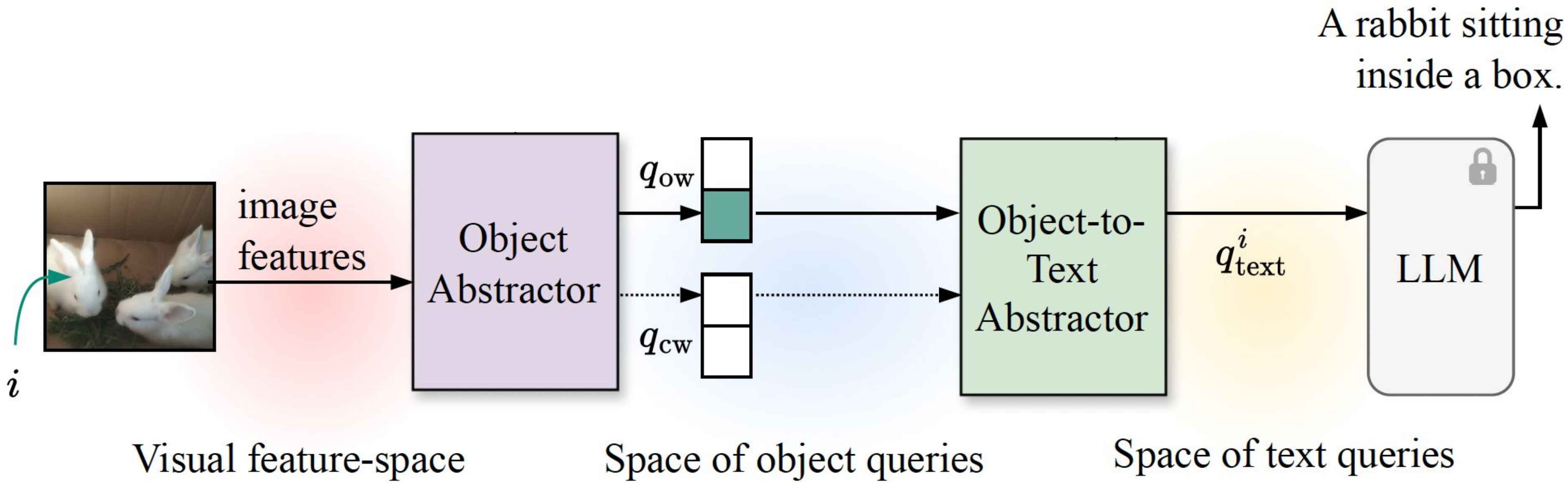
Abstractors for OW-VISCap: Challenges

- Haven’t been explored to connect object and language spaces
- How to extend them to the open-world without prompts?

OW-VISCapTor

OW-VISCapTor

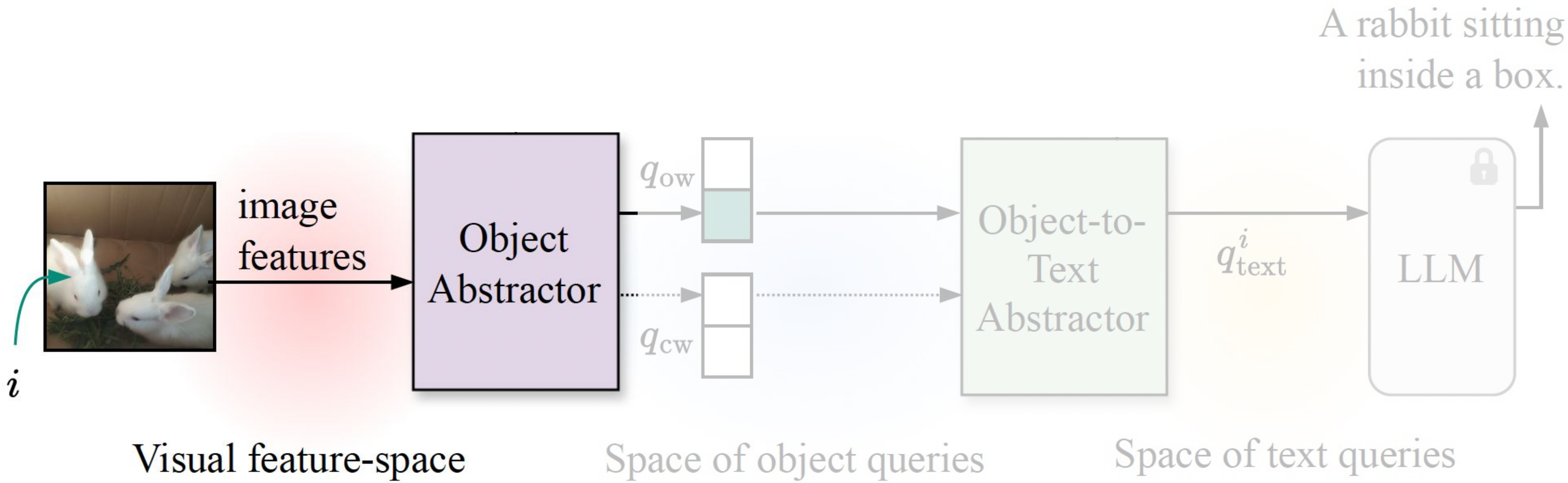
q_{ow} : open-world object queries
q_{cw} : closed-word object queries
q^i_{text} : text query for i-th object



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

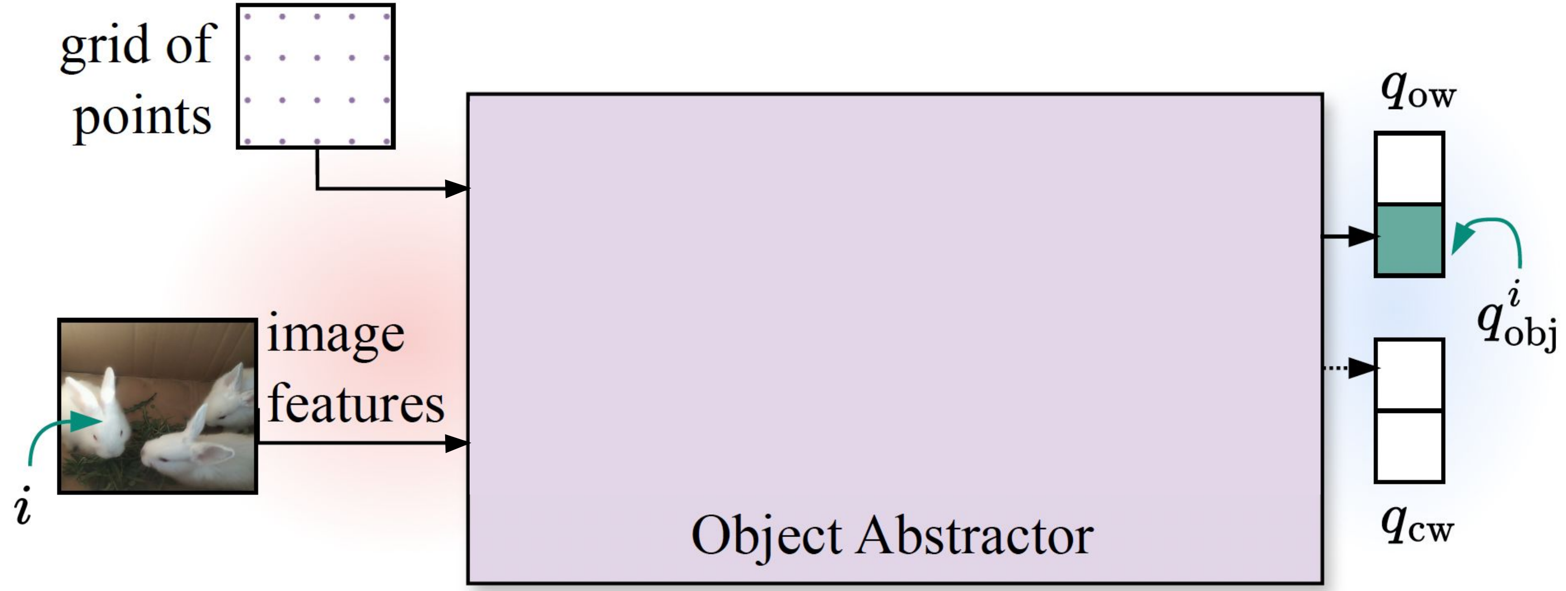
Open-World Object Discovery

q_{ow} : open-world object queries
q_{cw} : closed-word object queries
q^i_{text} : text query for i-th object



Open-World Object Discovery

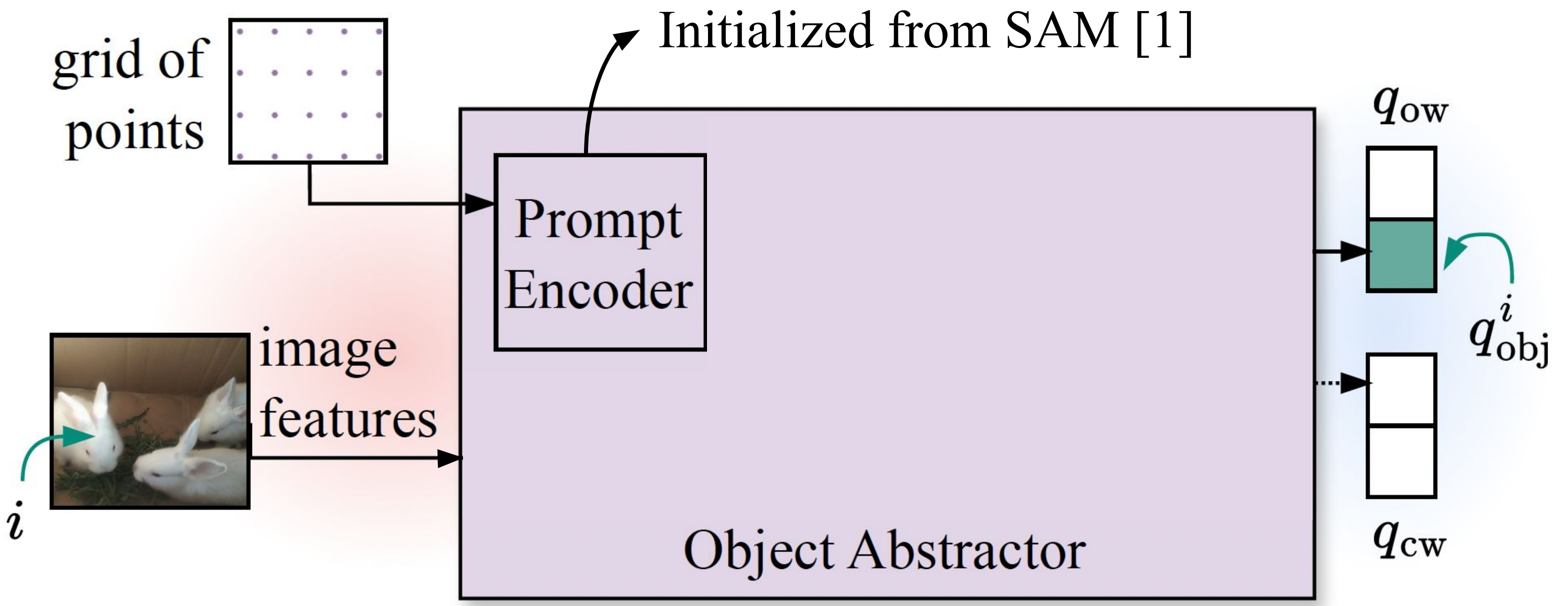
q_{ow} : open-world object queries
q_{cw} : closed-word object queries
q^i_{obj} : i-th object query



[1] Kirillov et al., arXiv 2023

Open-World Object Discovery

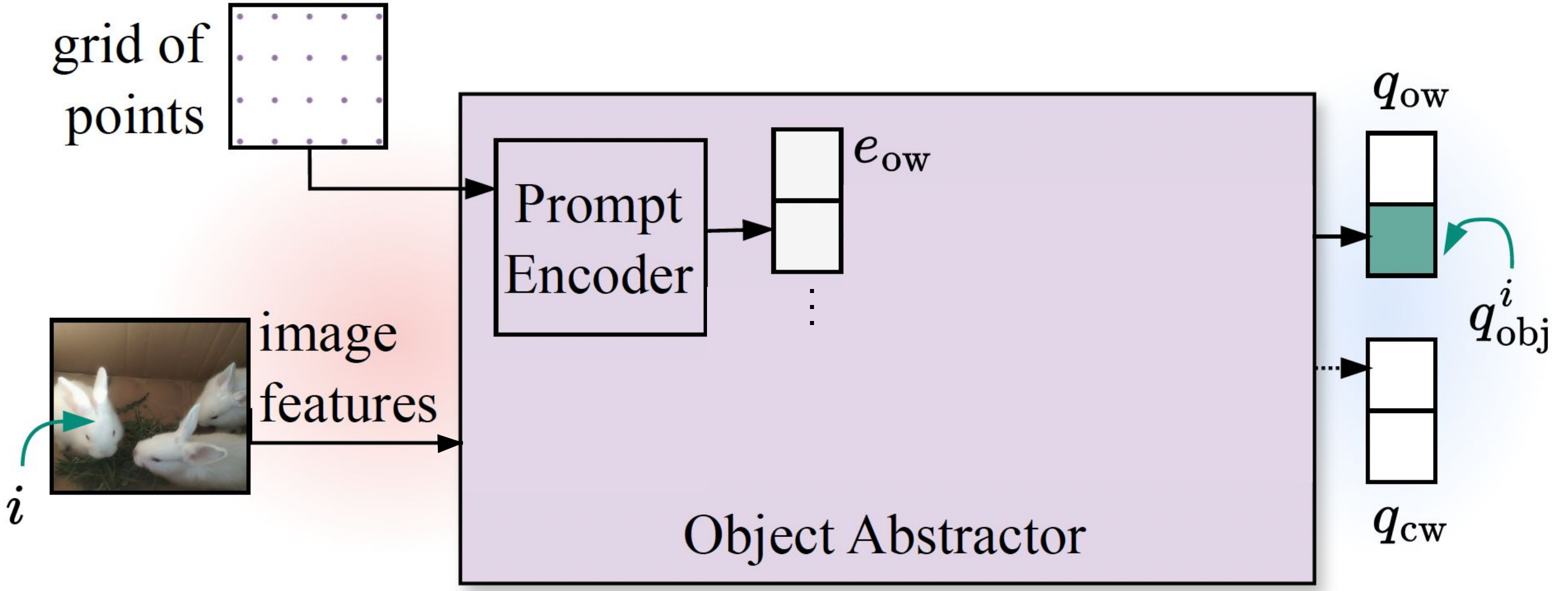
q_{ow} : open-world object queries
q_{cw} : closed-word object queries
q^i_{obj} : i-th object query



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

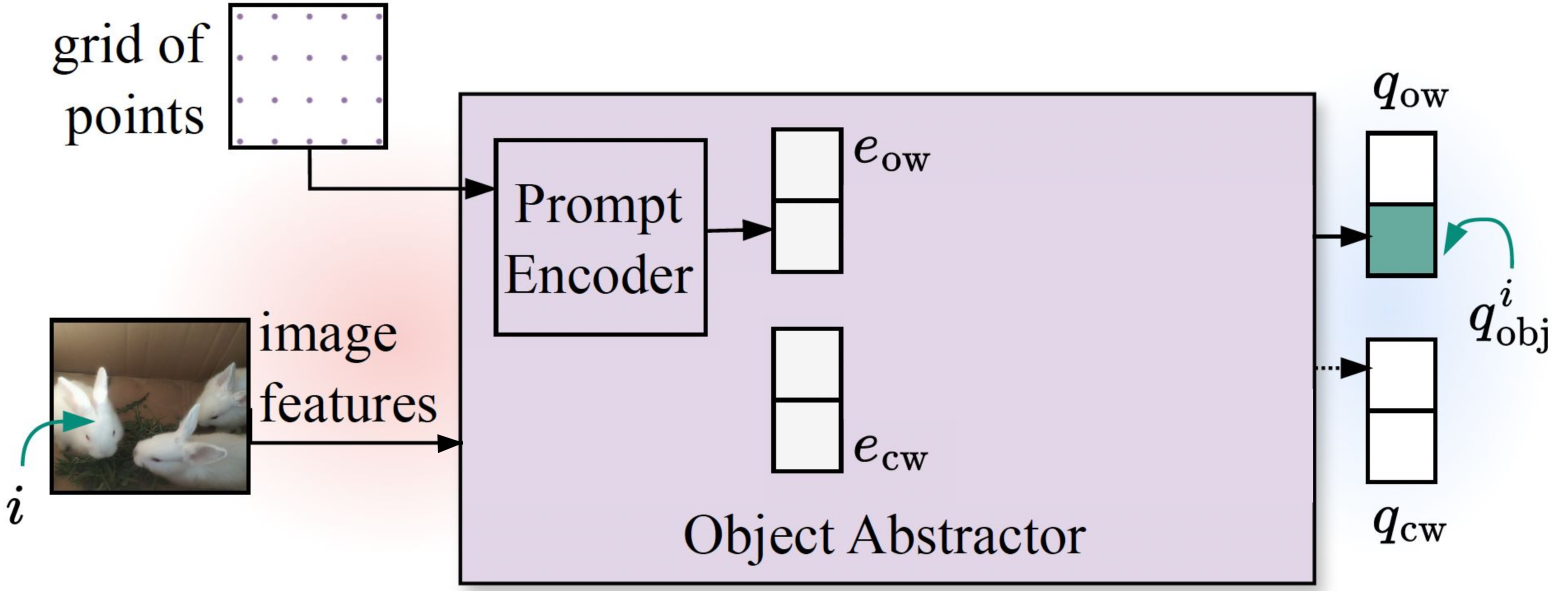
Open-World Object Discovery

e_{ow} : open-world embeddings
q_{ow} : open-world object queries
q_{cw} : closed-word object queries
q^i_{obj} : i-th object query



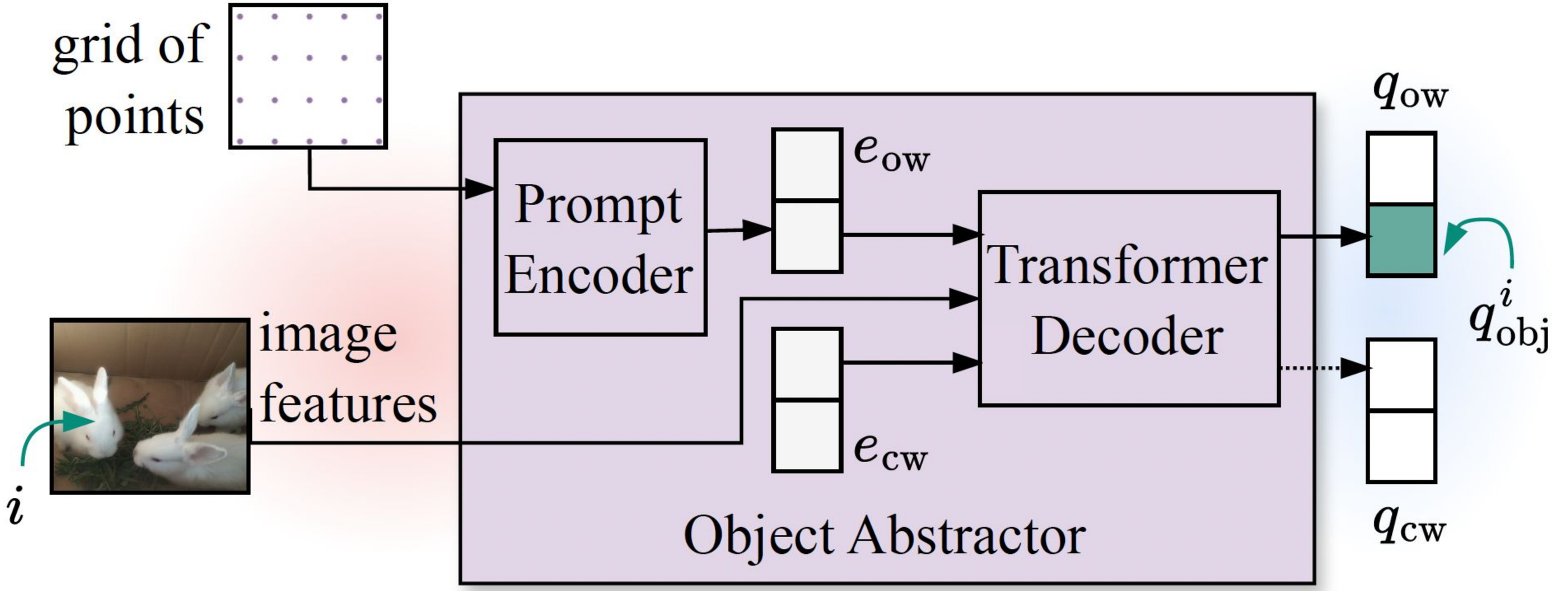
Open-World Object Discovery

e_{ow} : open-world embeddings
q_{ow} : open-world object queries
e_{cw} : closed-world embeddings
q_{cw} : closed-word object queries
q^i_{obj} : i-th object query



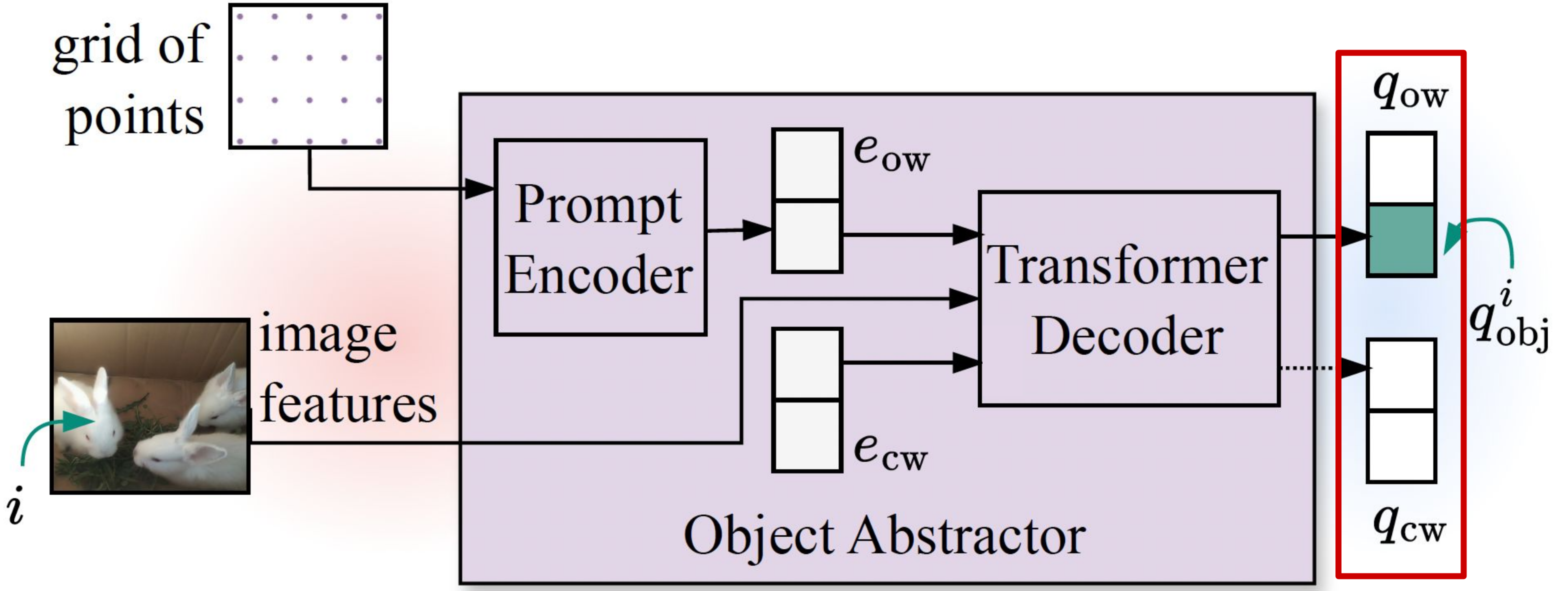
Open-World Object Discovery

e_{ow} : open-world embeddings
q_{ow} : open-world object queries
e_{cw} : closed-world embeddings
q_{cw} : closed-word object queries
q^{i}_{obj} : i-th object query



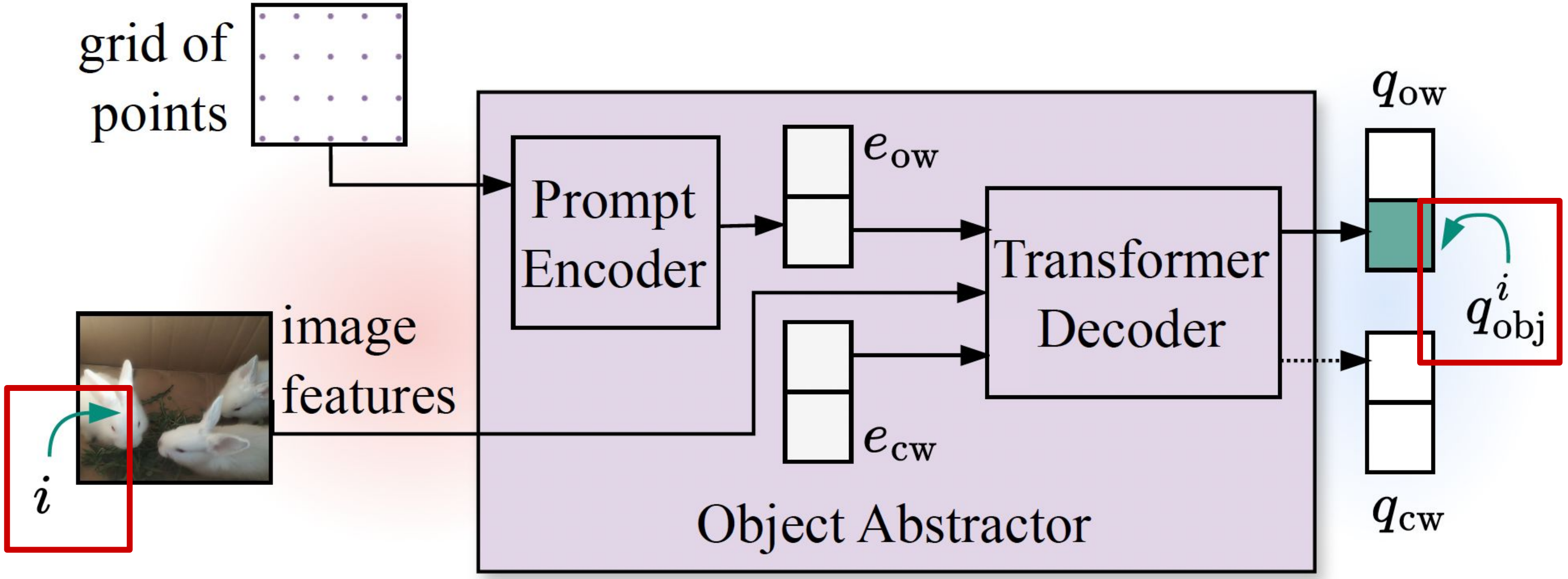
Open-World Object Discovery

e_{ow} : open-world embeddings
q_{ow} : open-world object queries
e_{cw} : closed-world embeddings
q_{cw} : closed-word object queries
q^i_{obj} : i-th object query



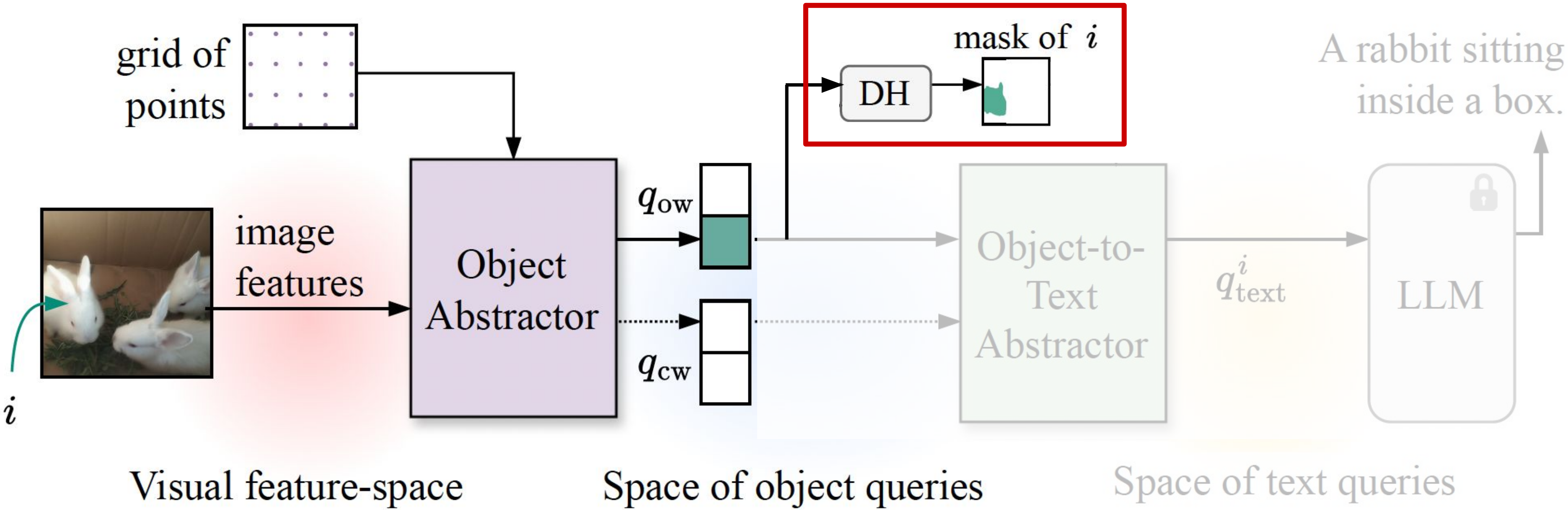
Open-World Object Discovery

e_{ow} : open-world embeddings
q_{ow} : open-world object queries
e_{cw} : closed-world embeddings
q_{cw} : closed-word object queries
q^i_{obj} : i-th object query



Object-Centric Captioning

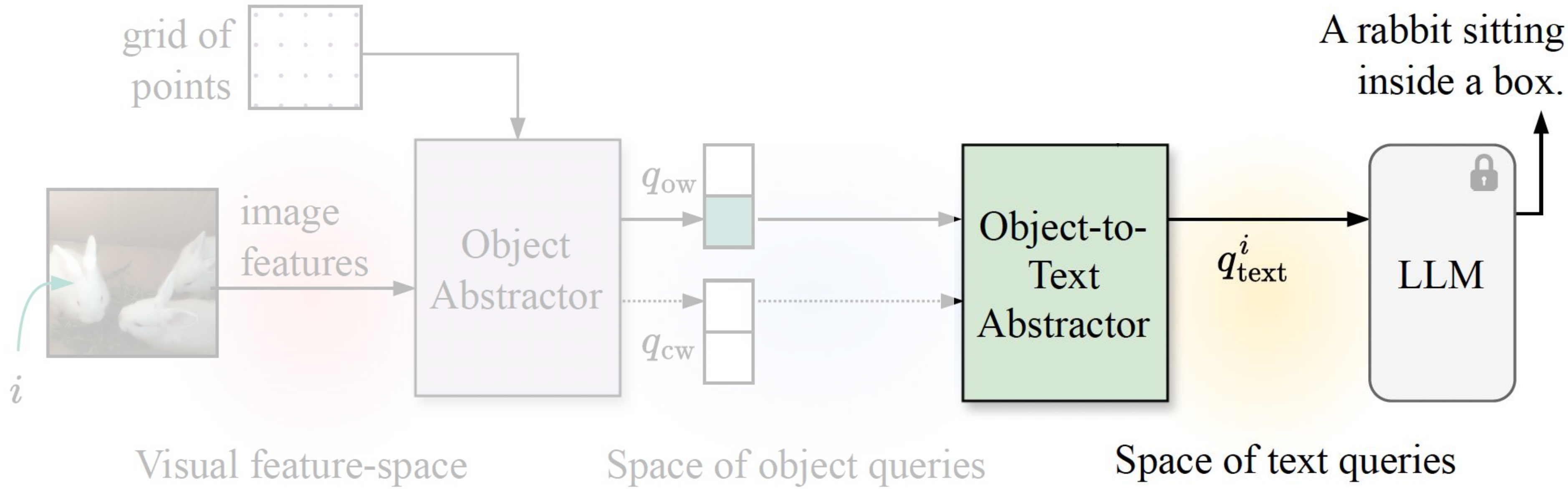
q_{ow} : open-world object queries
q_{cw} : closed-word object queries
q^i_{text} : text query for i-th object
DH : Detection Head



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

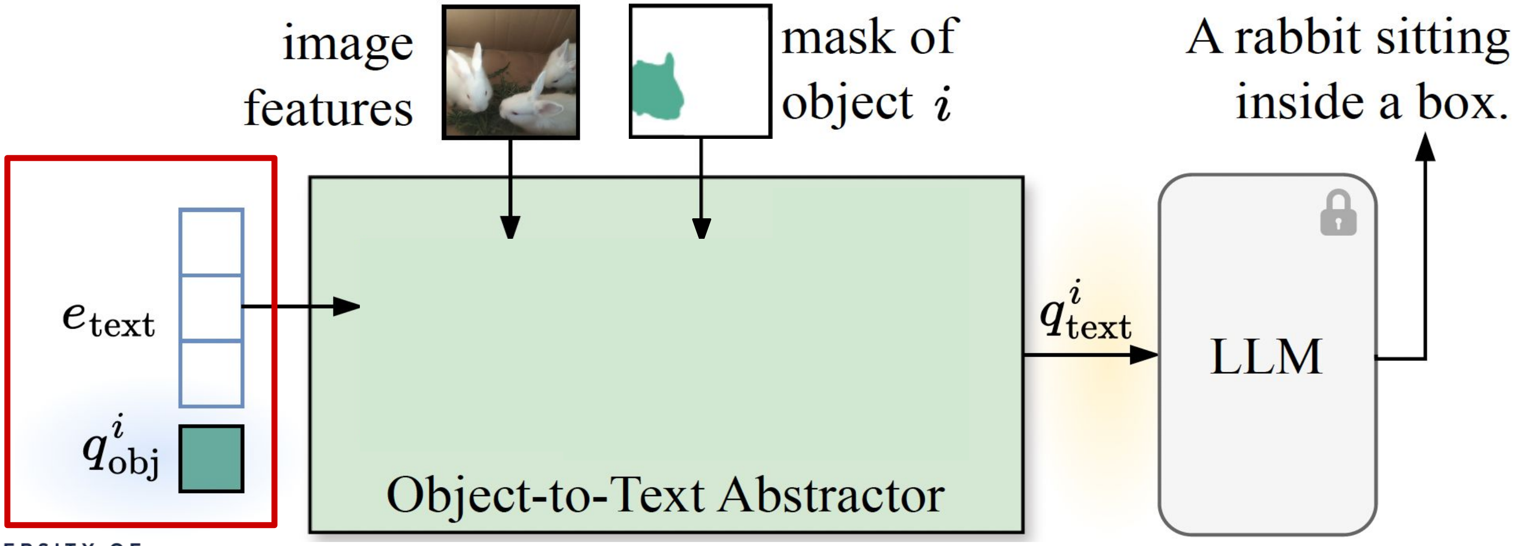
Object-Centric Captioning

q_{ow} : open-world object queries
q_{cw} : closed-word object queries
q^{i}_{text} : text query for i-th object



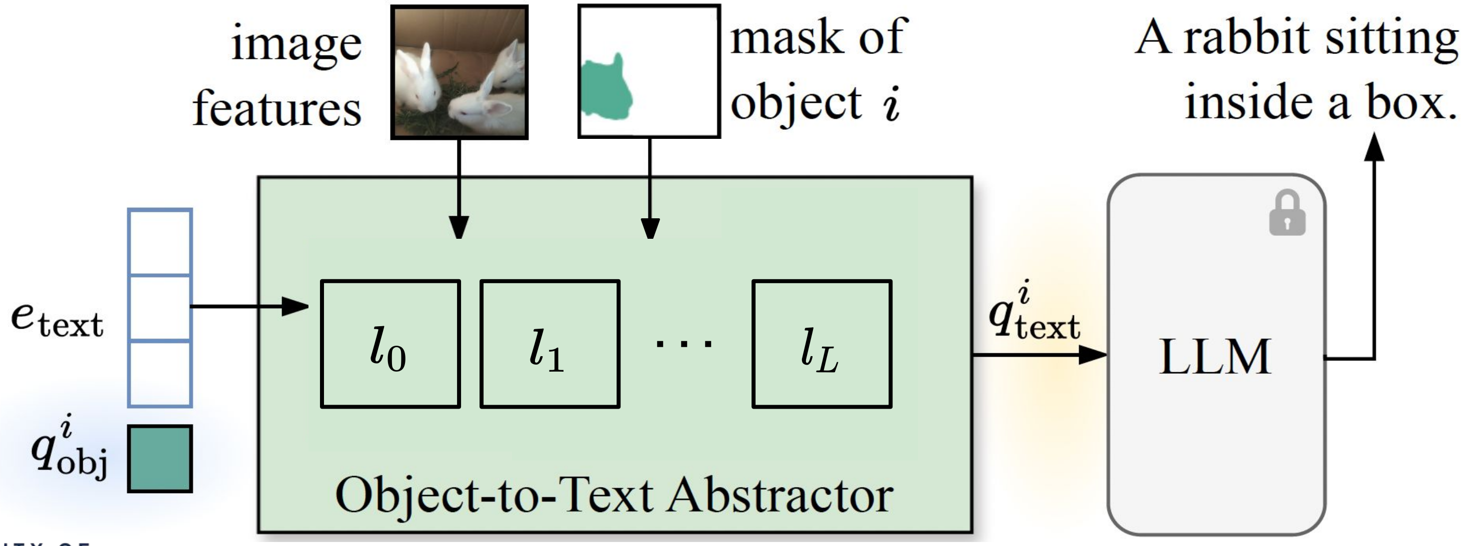
Object-Centric Captioning

e_{text} : text embeddings
q^i_{obj} : i-th object query
q^i_{text} : text query for i-th object



Object-Centric Captioning

e_{text} : text embeddings
q^i_{obj} : i-th object query
q^i_{text} : text query for i-th object



Results

Results on BURST [1] Dataset

Segmentation of open-world and closed-world objects

[1] Athar et al., WACV 2023
[2] Liu et al.' CVPR 2022
[3] Cheng et al., CVPR 2022
[4] Cheng et al., Neurips 2021
[5] Cheng et al., ICCV 2023
[6] Qi et al., PAMI 2022

Method	**Accuracy**		
	Unseen	Overall	Seen
OWTB [2]	38.8	55.8	59.8
Mask2Former [3] + STCN [4]	25.0	64.6	71.0
Mask2Former [3] + DEVA [5]	42.3	**69.5**	**74.6**
EntitySeg [6] + DEVA [5]	<u>49.6</u>	68.8	72.7
Ours + DEVA [5]	**55.2**	<u>69.0</u>	<u>73.5</u>

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Results on VidSTG [1] Dataset

[1] Zhang et al., CVPR 2020
[2] Zhou arXiv 2023
[3] Choudhuri et al., CVPR 2023

Bounding box detections and captioning on closed-world objects

Method	Mode	Captioning accuracy	Overall accuracy
DenseVOC-DS (joint training) [2]	offline	36.8	51.6
DenseVOC-DS (disjoint training) [2]	offline	10.0	28.0
Ours + CAROQ [3]	online	**43.9**	**53.1**

<u>a large construction truck with a trailer on it.</u>

a car is driving in the rain on a street.

...

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

a large construction truck with a trailer on it.

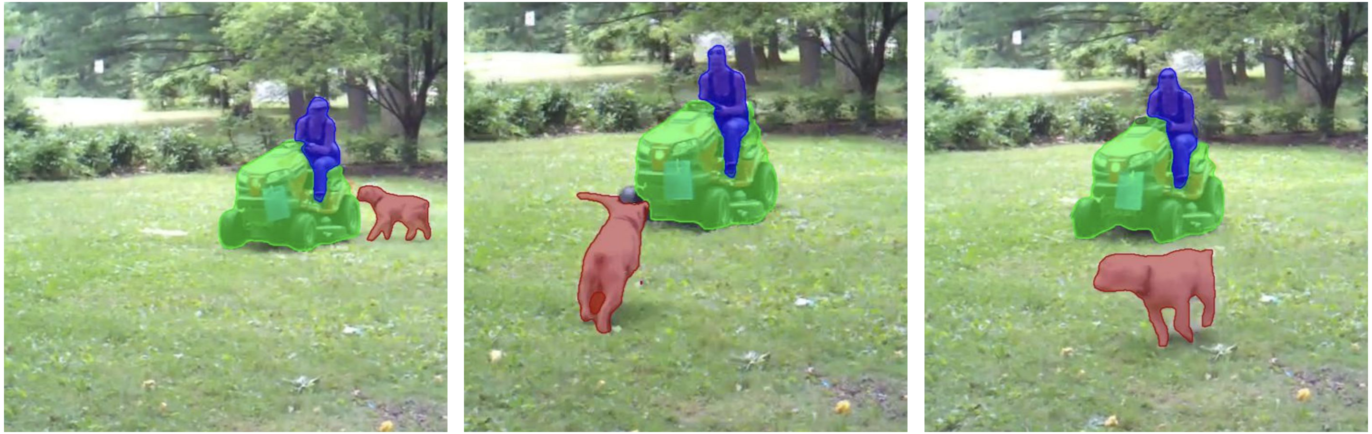
a car is driving in the rain on a street.

...

a tractor with black and orange front and rear.

a woman is riding an orange lawn mower.

a white dog near a tractor.



UNIVERSITY OF ILLINOIS
URBANA-CHAMPAIGN

To Summarize

- We propose a new task: Open-World Video Instance Segmentation and Captioning (OW-VISCap).

To Summarize

- We propose a new task: Open-World Video Instance Segmentation and Captioning (OW-VISCap).
- OW-VISCapTor:
  - **Object abstractor**: spatially rich open-world object queries
  - **Object-to-text abstractor**: rich object-centric captions

To Summarize

- We propose a new task: Open-World Video Instance Segmentation and Captioning (OW-VISCap).
- OW-VISCapTor:
  - **Object abstractor**: spatially rich open-world object queries
  - **Object-to-text abstractor**: rich object-centric captions
- Our generalized approach surpasses individual SOTA on open-world object discovery and video object captioning

Thank You!

Please visit our poster on

Wed, Dec 11, Poster Session 2 (4:30 - 7:30 p.m. PST)

Website:
https://anwesachoudhuri.github.io/OpenWorldVISCap/

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN