JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Paper Tag: THU-AM-283

Zero-shot Referring Image Segmentation with Global-Local Context Features



Seonghoon Yu[1] Paul Hongsuck Seo[2] Jeany Son[1]

[1]AI Graduate School, GIST [2]Google Research

Quick Preview:

Referring Image Segmentation (RIS)

Input image

Output mask

Input text

“a cat is lying on the seat of the scooter”

Quick Preview:

Referring Image Segmentation (RIS)

Input image

Output mask

Input text

“a cat is lying on the seat of the scooter”

Problems

Labelling RIS dataset is costly

1

“a scooter
with two cats sitting on”

2

Quick Preview:

Referring Image Segmentation (RIS)

Input image

Output mask

Input text

“a cat is lying on the seat of the scooter”

Problems

Labelling RIS dataset is costly

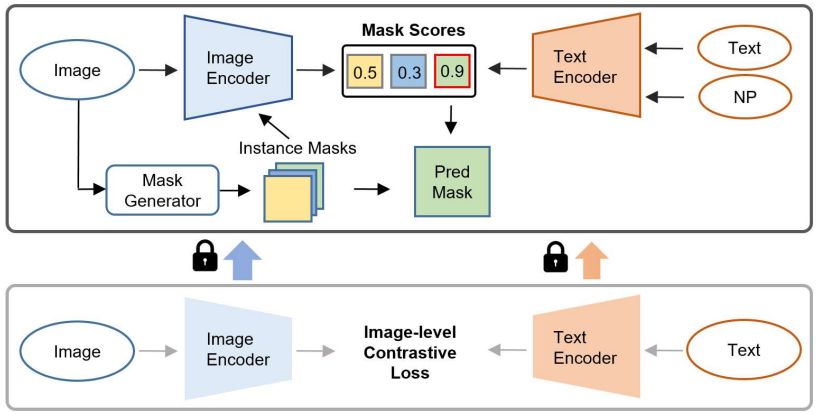
1

“a scooter
with two cats sitting on”

2

New task of Zero-shot RIS

Zero-shot RIS without any additional training



Pre-trained Vision-language Model (CLIP)

Quick Preview:

Referring Image Segmentation (RIS)

Input image

Output mask

Input text

“a cat is lying on the seat of the scooter”

New task of Zero-shot RIS

Zero-shot RIS without any additional training

Pre-trained Vision-language Model (CLIP)

Problems

Labelling RIS dataset is costly

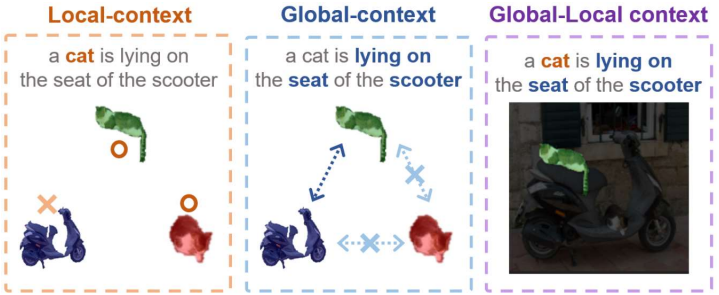
1

“a scooter
with two cats sitting on”

2

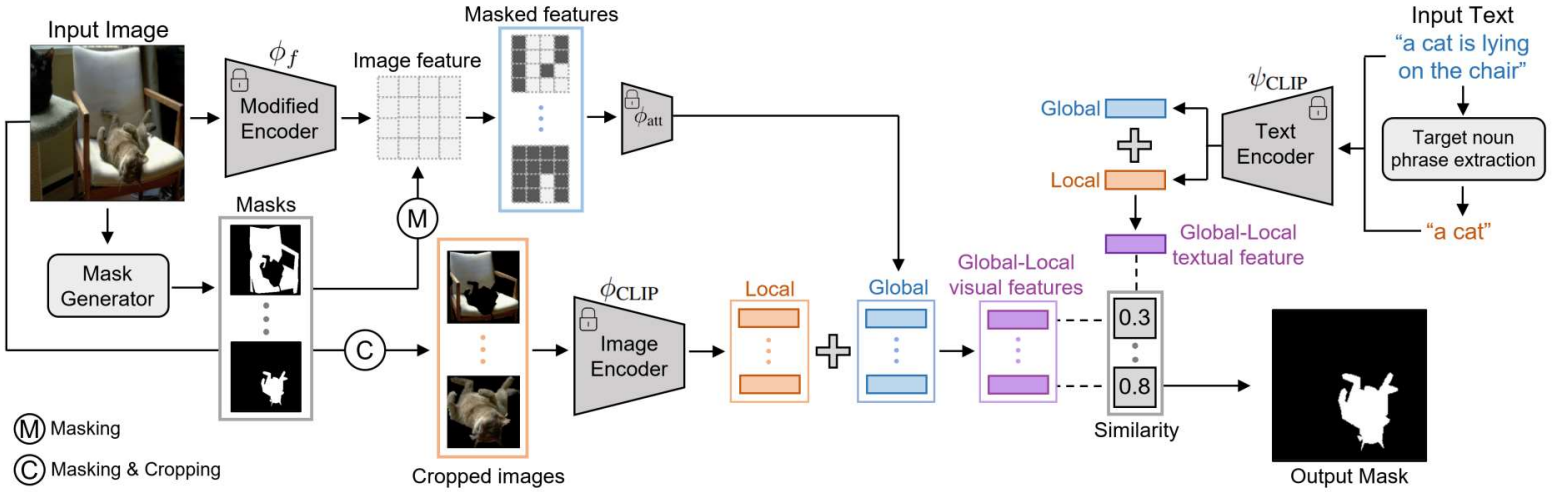
Motivation

Global-Local Context in RIS



Quick Preview:

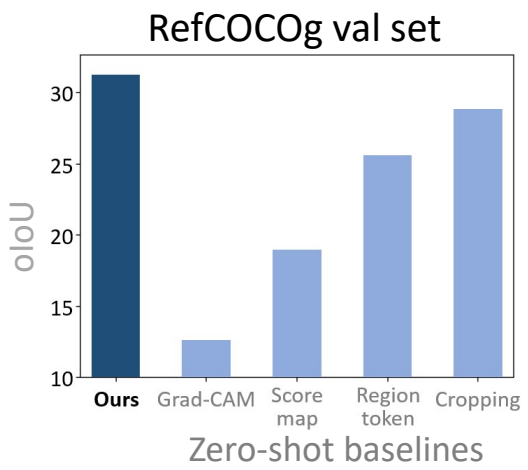
Overall Framework

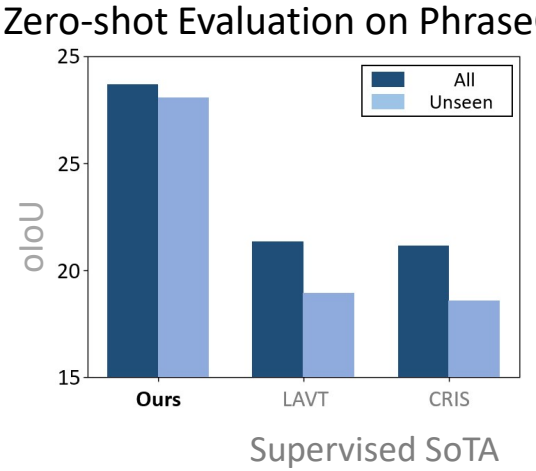


Quick Preview:

Overall Framework

Experiments





Details

Referring Image Segmentation

Input image

Input text

“a cat is lying on the seat of the scooter”

Referring Image Segmentation

Input image

Output mask

Input text

“a cat is lying on the seat of the scooter”

Referring Image Segmentation

Input image

Input text

“the bottom cat”

Referring Image Segmentation

Input image

Output mask

Input text

“the bottom cat”

Referring Image Segmentation

Input image

Input text

“a scooter with two cats sitting on”

Referring Image Segmentation

Input image

Output mask

Input text

“a scooter with two cats sitting on”

Challenges

1. High-level understanding of language

"a cat is lying on the seat of the scooter"

2. Comprehensive understanding of an image

3. Dense instance-level prediction

4. Select only one instance
from several objects of the same class

Challenges

1. High-level understanding of language

"a cat is lying on the seat of the scooter"

2. Comprehensive understanding of an image

3. Dense instance-level prediction

4. Select only one instance
from several objects of the same class

Challenges

1. High-level understanding of language

"a cat is lying on the seat of the scooter"

2. Comprehensive understanding of an image

3. Dense instance-level prediction

4. Select only one instance
from several objects of the same class

Challenges

1. High-level understanding of language

"a cat is lying on the seat of the scooter"

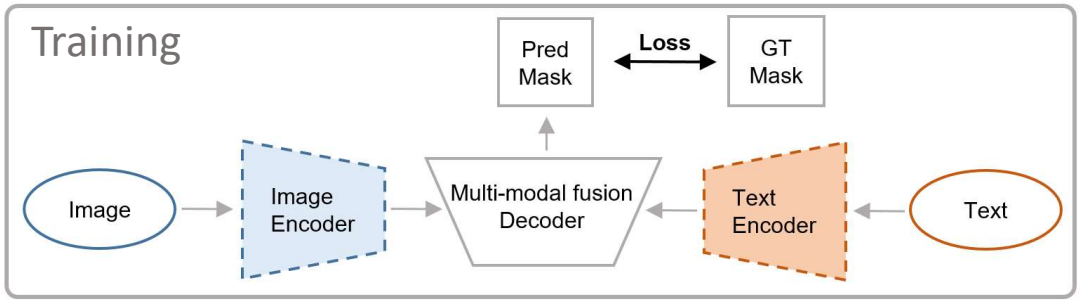
2. Comprehensive understanding of an image

3. Dense instance-level prediction

4. Select only one instance
from several objects of the same class

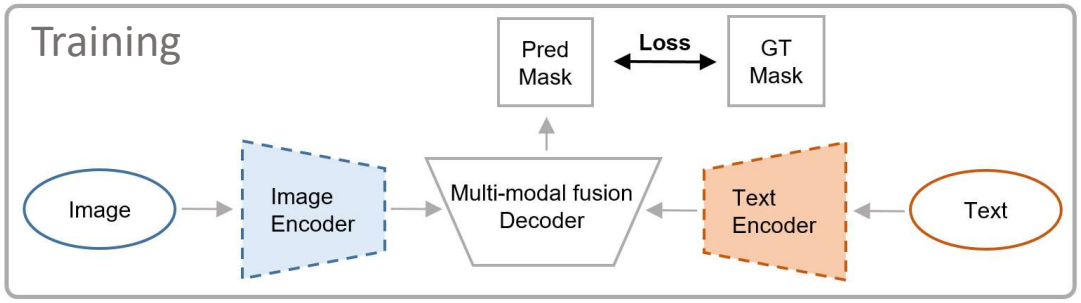
Problems

Fully supervised RIS



Problems

Fully supervised RIS



Labelling RIS dataset is costly

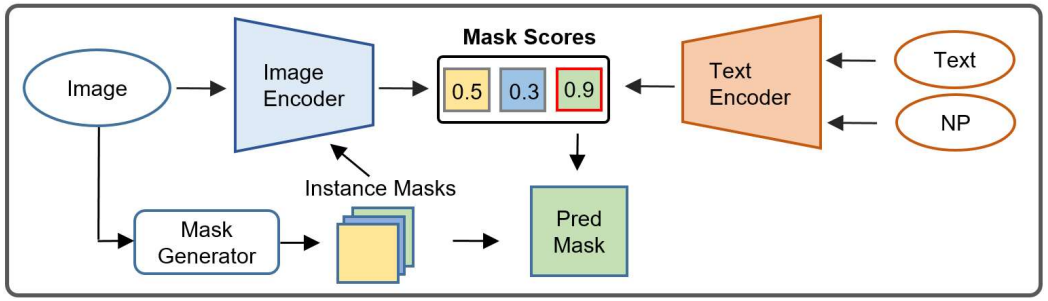
1

"a cat is lying on
the seat of the scooter"

2

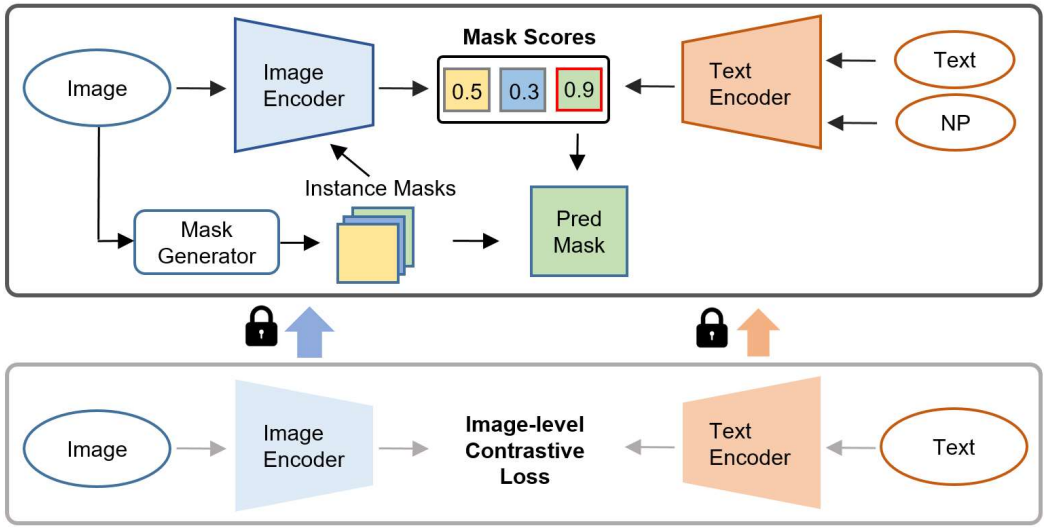
The new task of Zero-Shot RIS

Zero-shot RIS without any additional training



The new task of Zero-Shot RIS

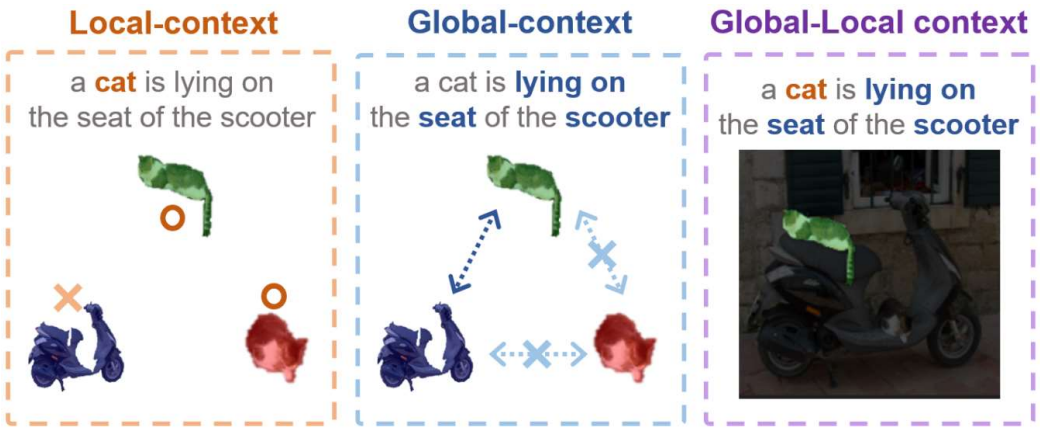
Zero-shot RIS without any additional training



Pre-trained Vision-language Model (CLIP)

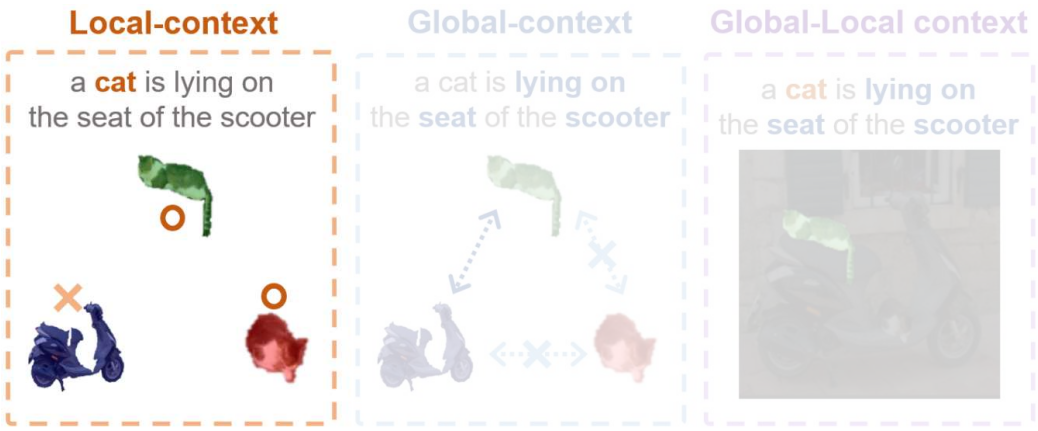
Motivation

Global-Local Context in RIS



- **Global-Local context:** consider jointly Local- and Global- context

Motivation

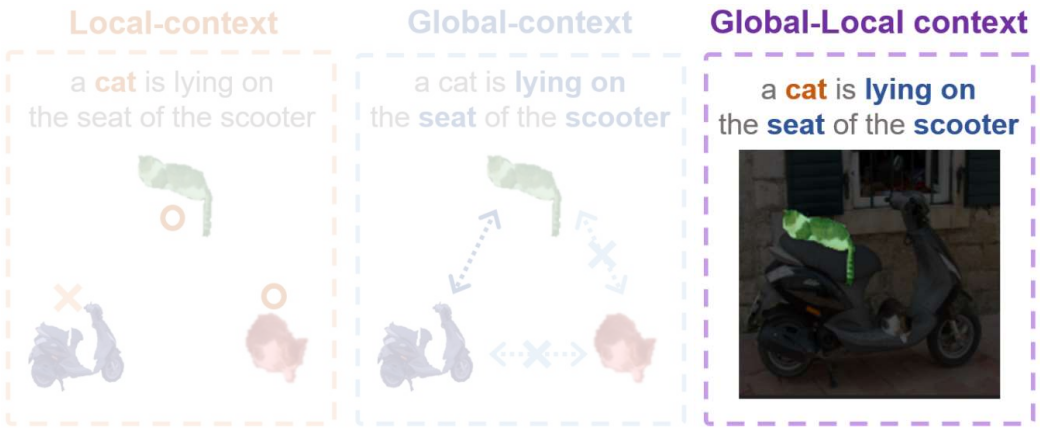


- **Local-context:** focus on the target class to be selected

Motivation

- **Global-context:** consider the relations between objects

Motivation

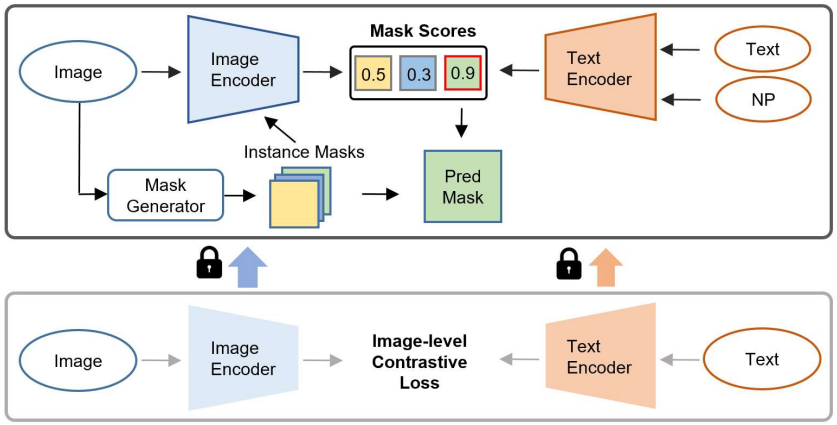


- Extract **Global-Local context** features on both visual and textual modalities

Contributions

1. New task of Zero-shot RIS based on CLIP without any additional training

1



2

3

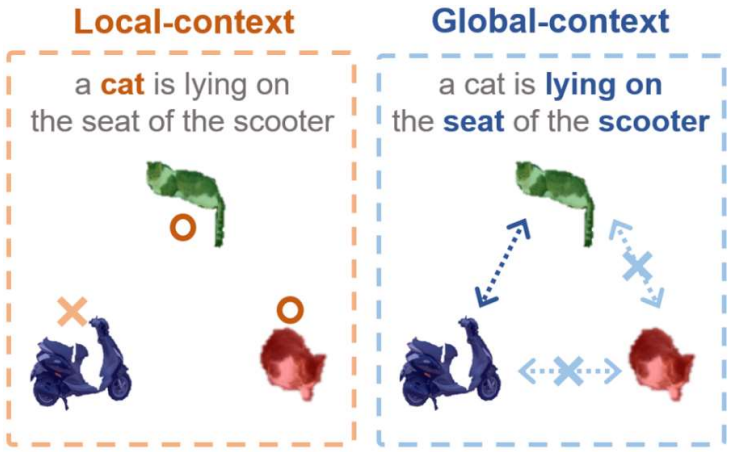
Contributions

1. New task of Zero-shot RIS based on CLIP without any additional training

2. Visual and textual encoders
   to integrate Global- and Local- context features of images and sentences, respectively

1

2



3

Contributions

1. New task of Zero-shot RIS based on CLIP without any additional training

2. Visual and textual encoders
   to integrate Global- and Local- context features of images and sentences, respectively

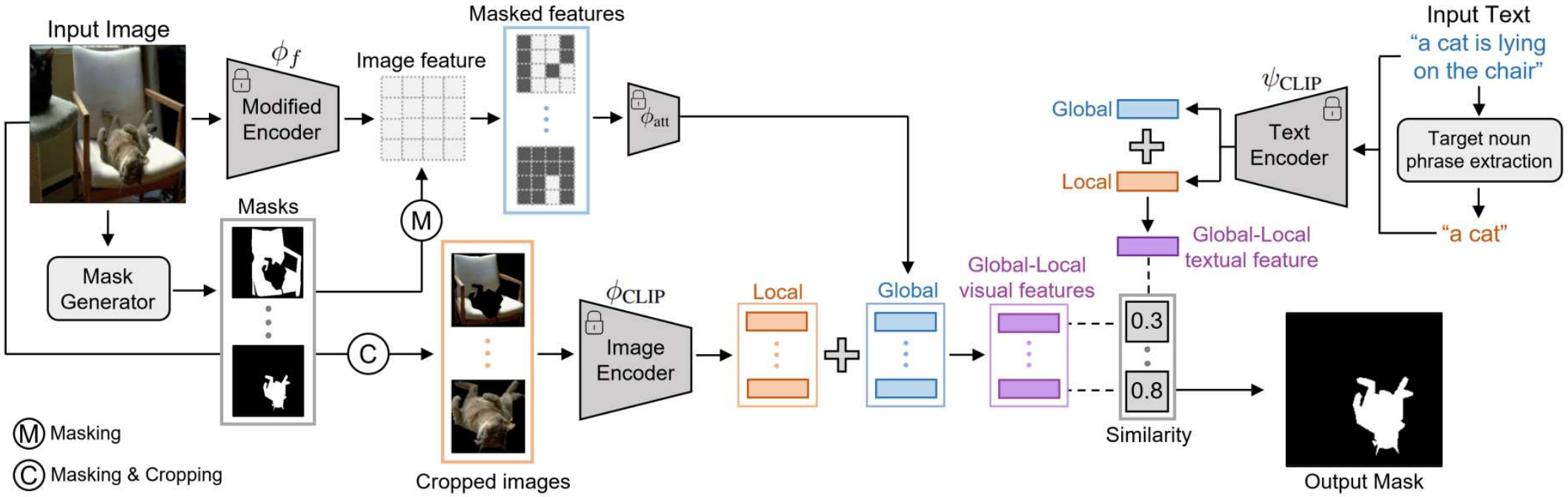
3. Global-Local context features
   to capture the target semantics as well as the relations between objects

1

2

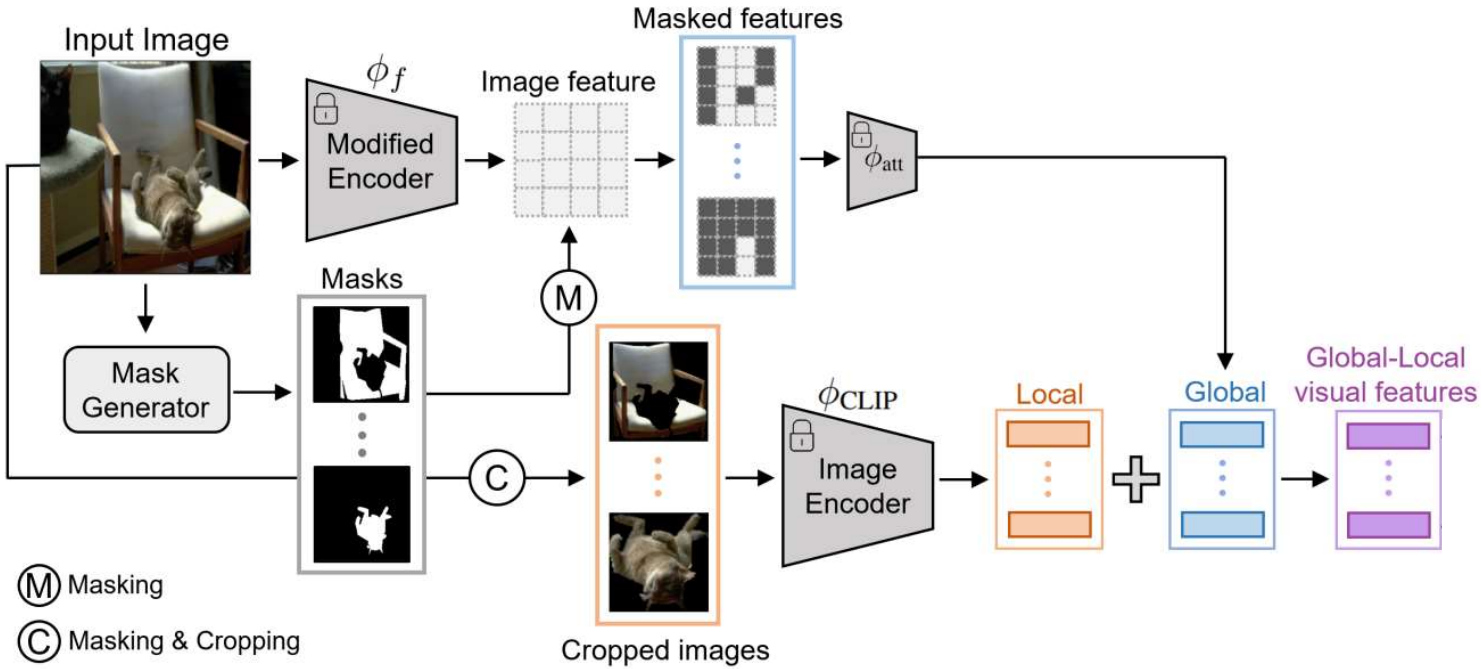
3

Overall Framework



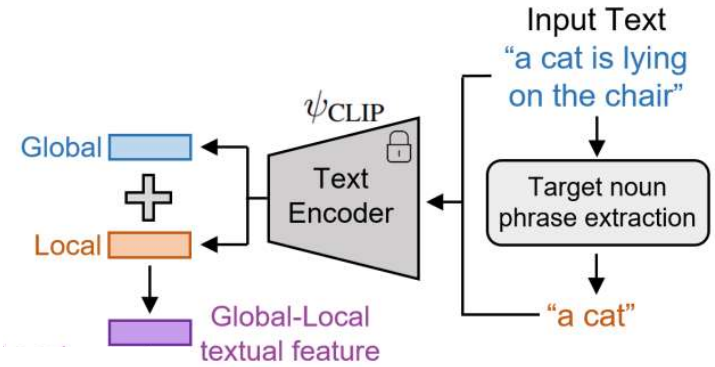
Overall Framework

1. Global-Local visual encoder



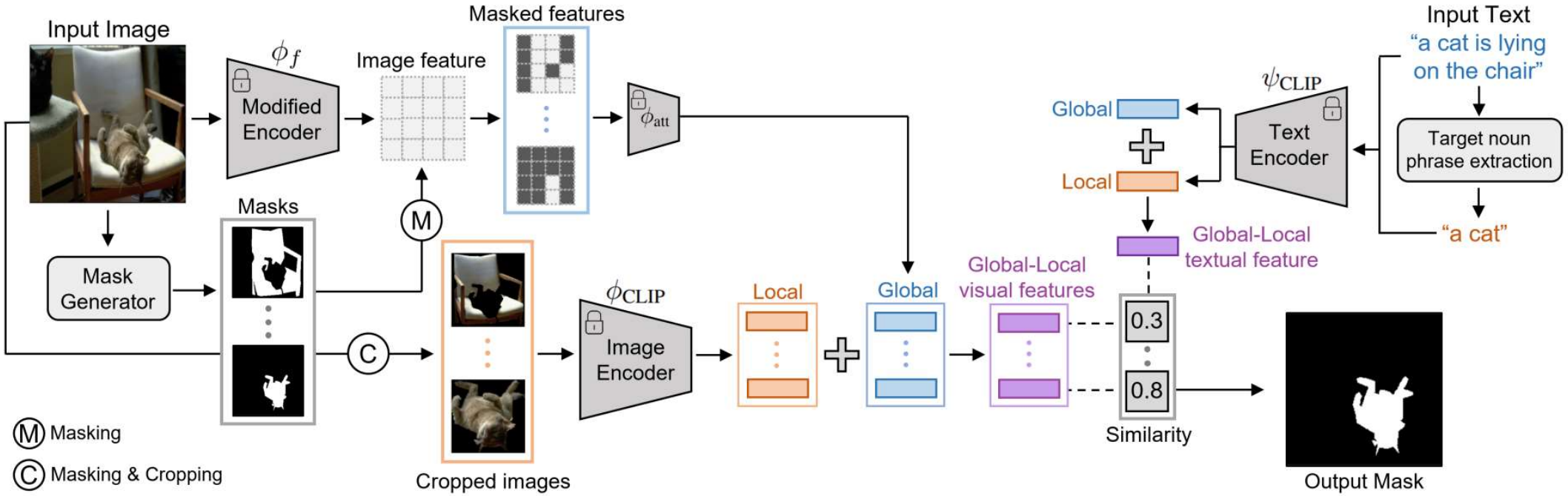
Overall Framework

2. Global-Local textual encoder



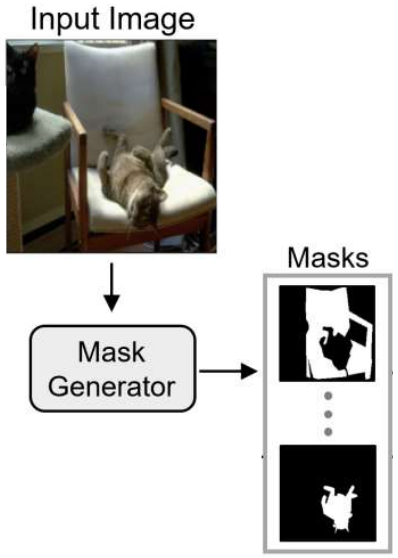
Overall Framework

How our Global-Local CLIP works

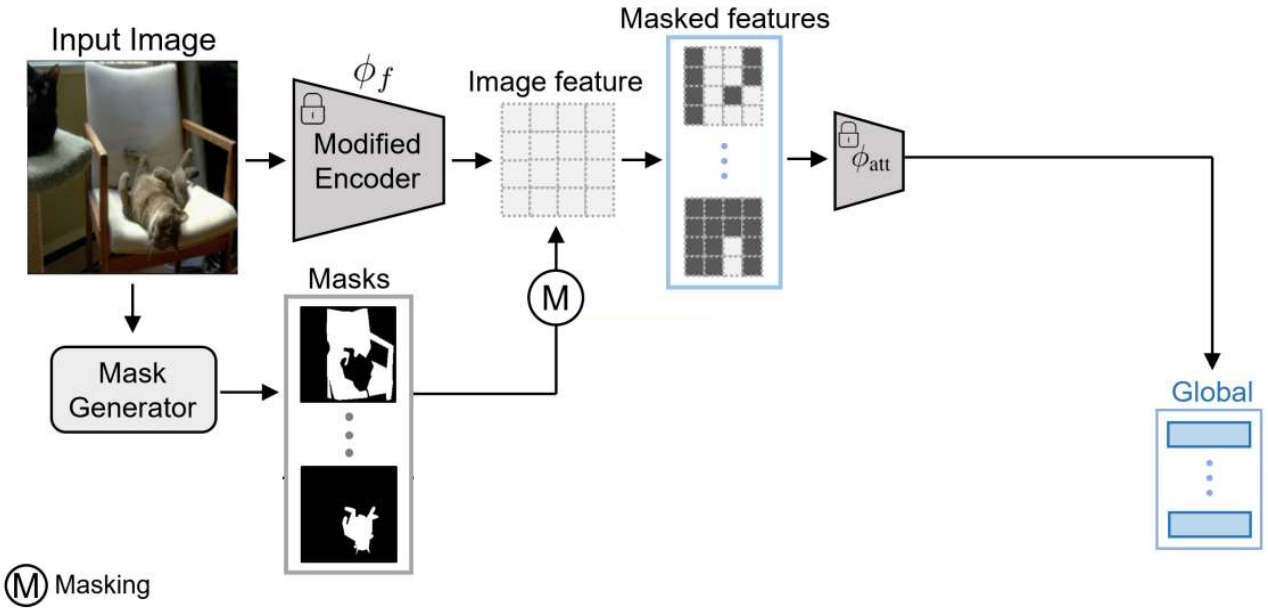


Overall Framework

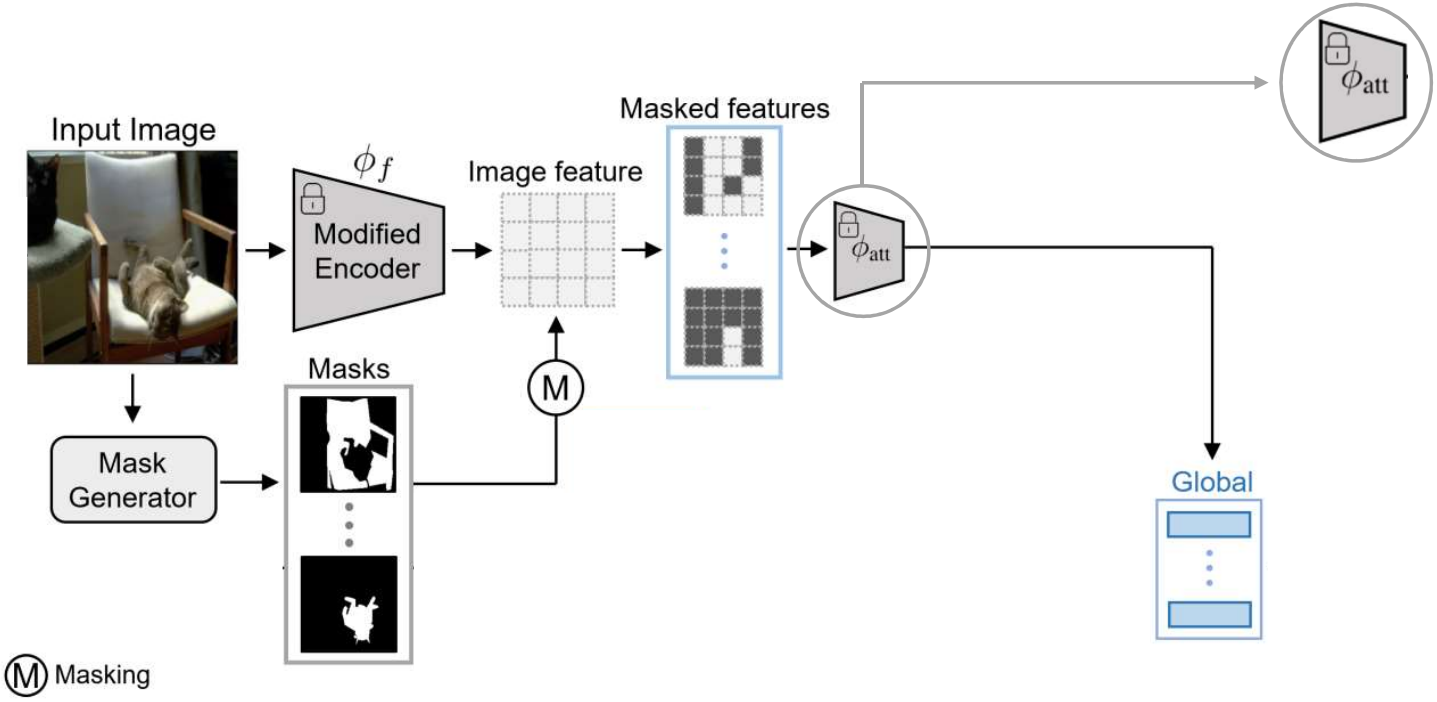
Mask proposals from the off-the-shelf instance segmentation model



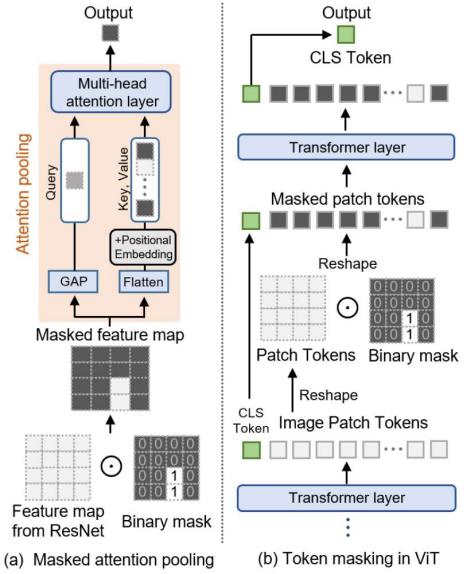
Global-context Visual Features



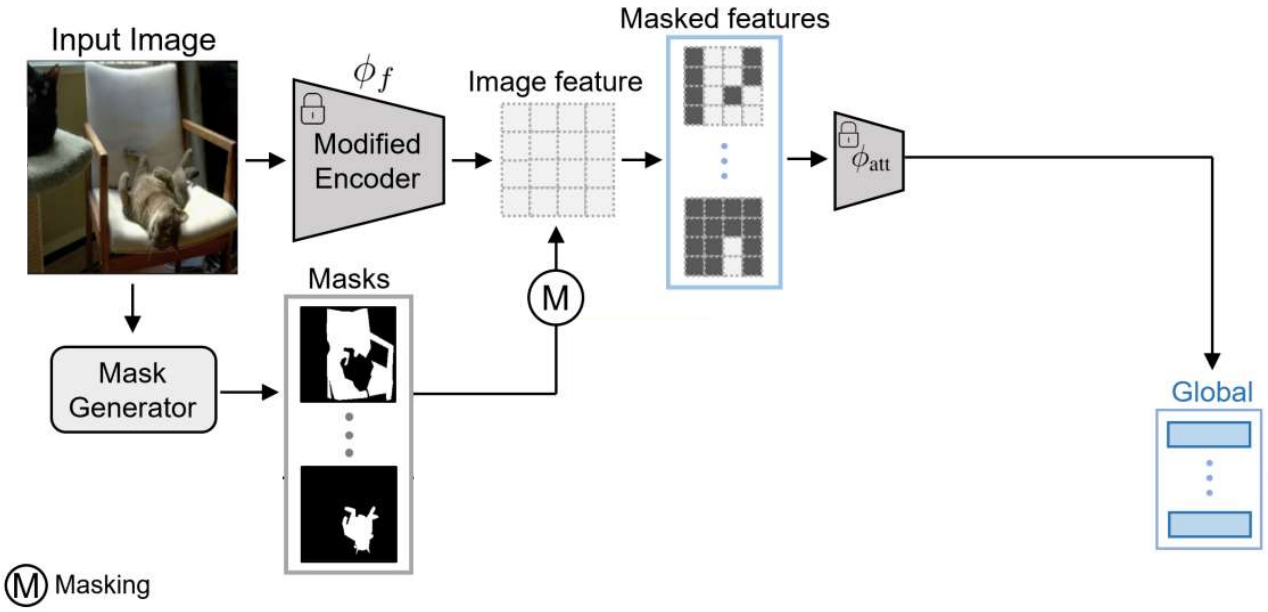
Global-context Visual Features



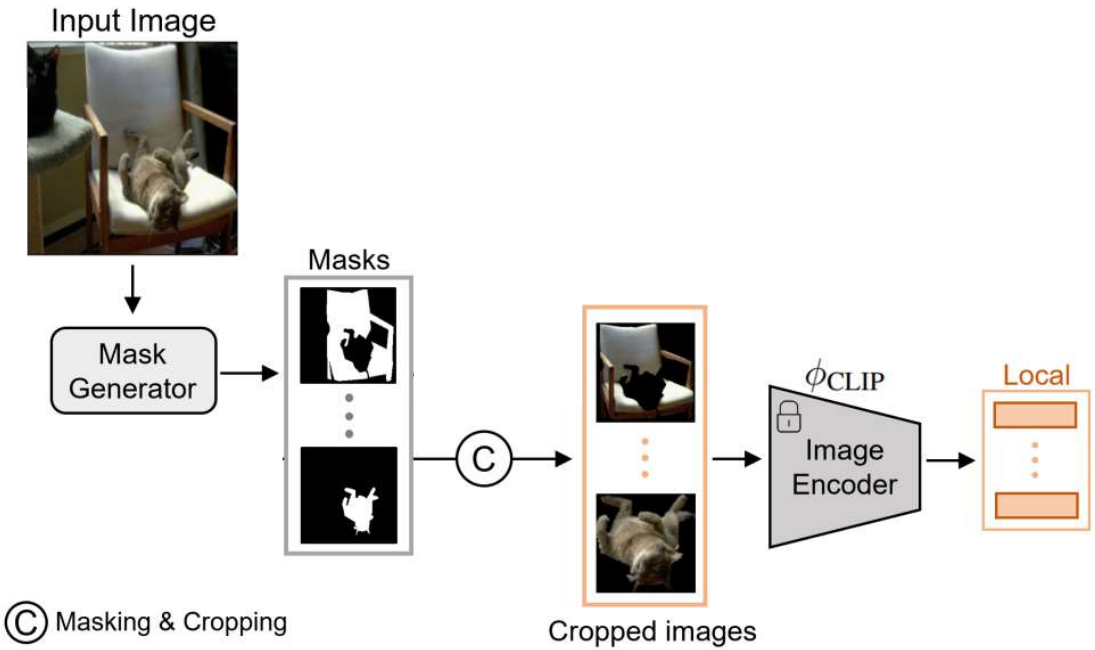
Mask-guided
global-context visual encoder



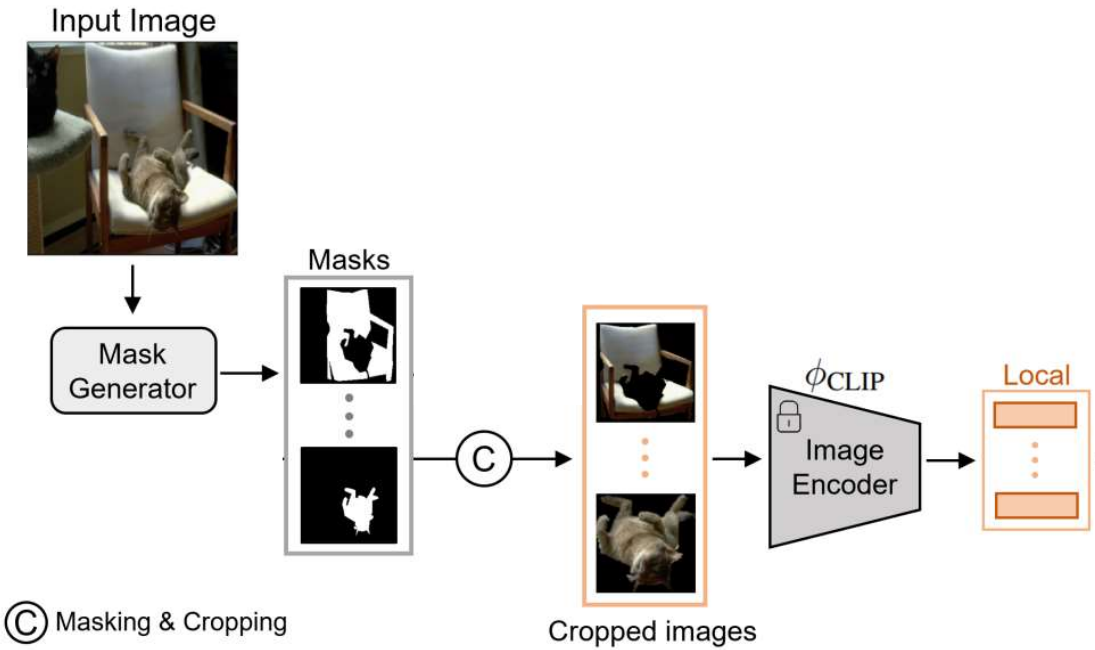
Global-context Visual Features

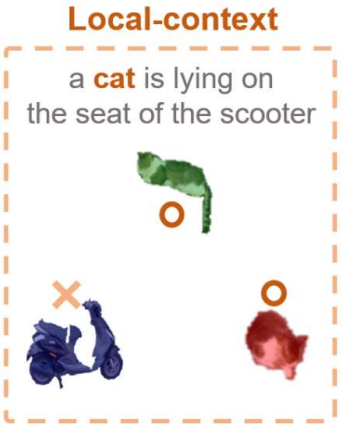


Local-context Visual Features

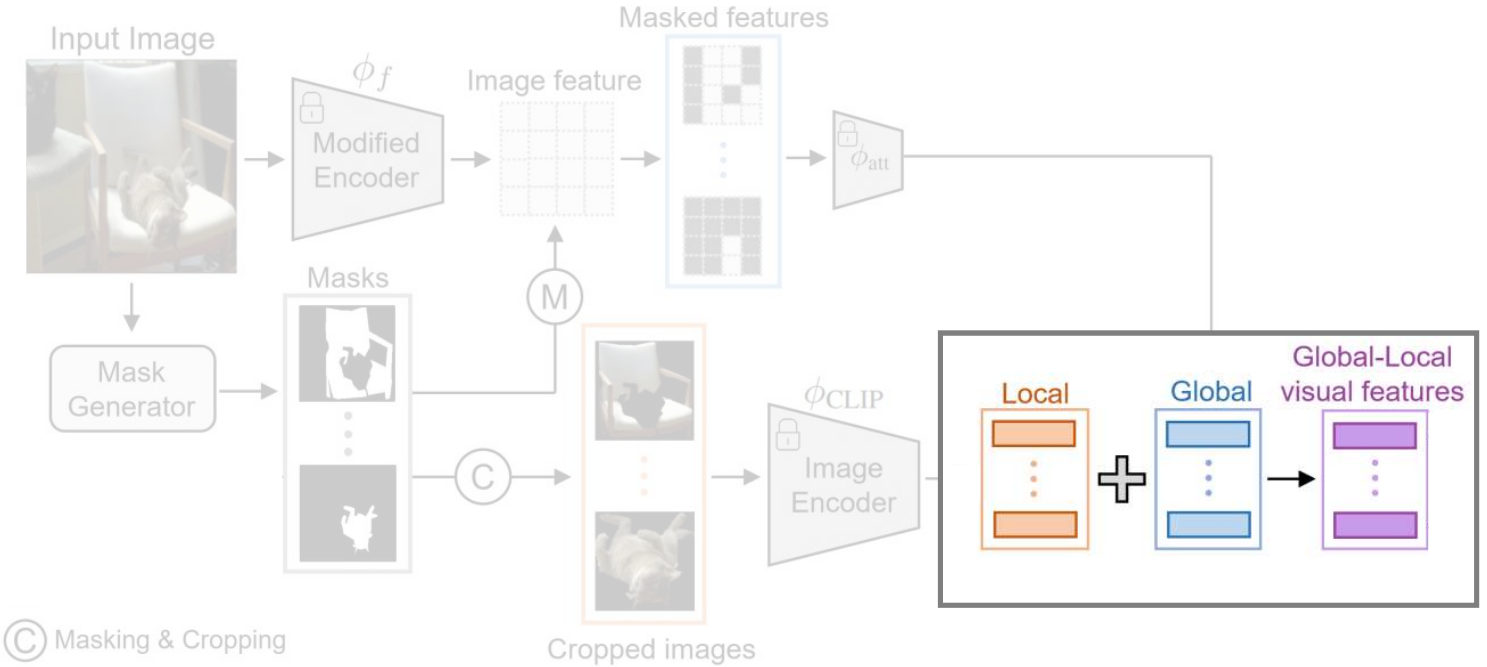


Local-context Visual Features





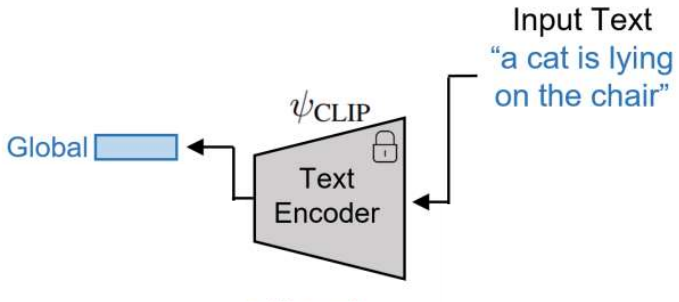
Global-Local Context Visual Features



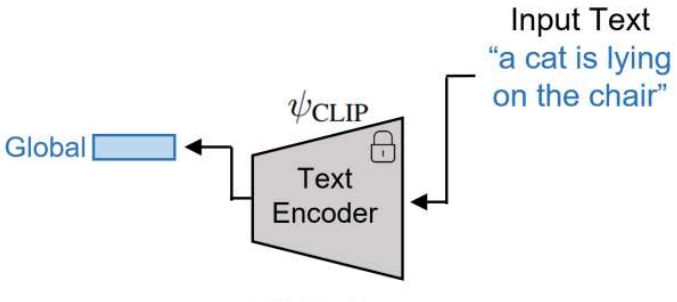
Global-Local context

a **cat** is **lying on** the **seat** of the **scooter**

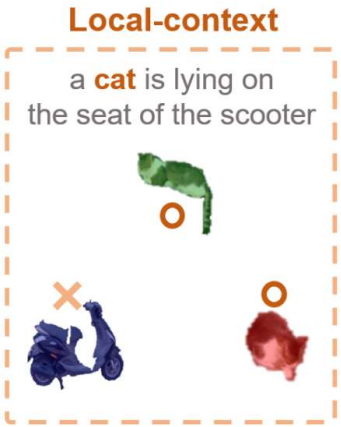
Global-context Textual Features

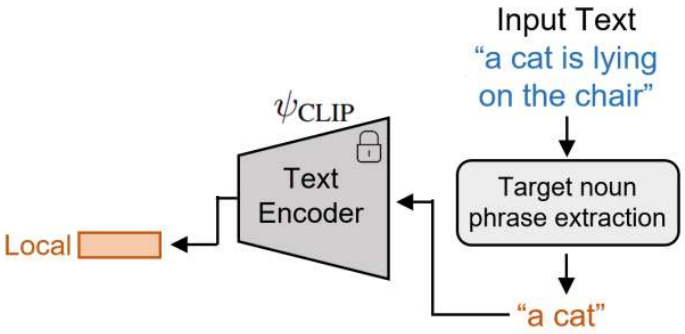


Global-context Textual Features

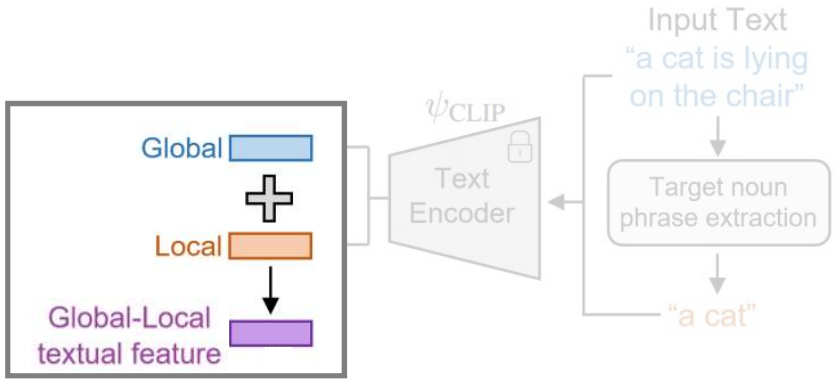


Local-context Textual Features

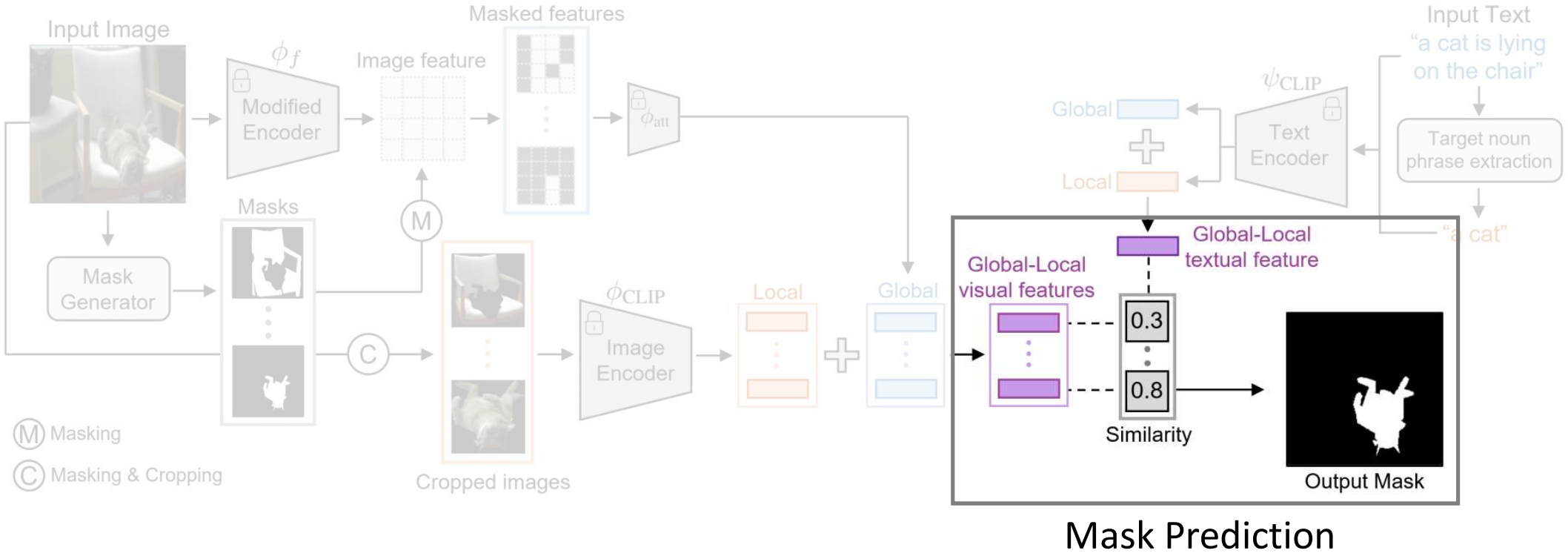




Global-Local context Textual Features

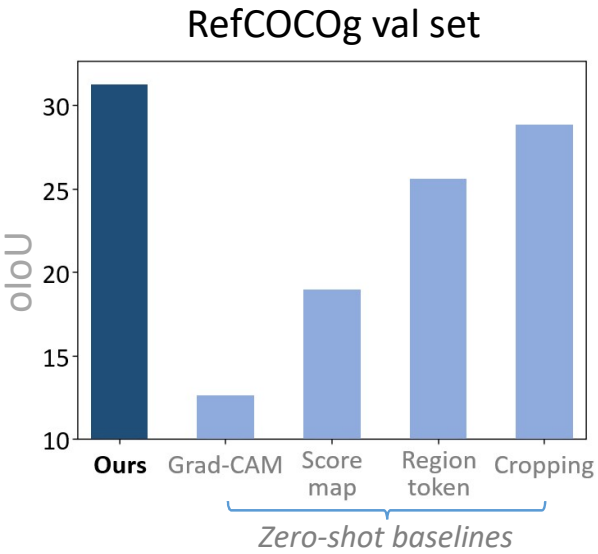


Overall Framework

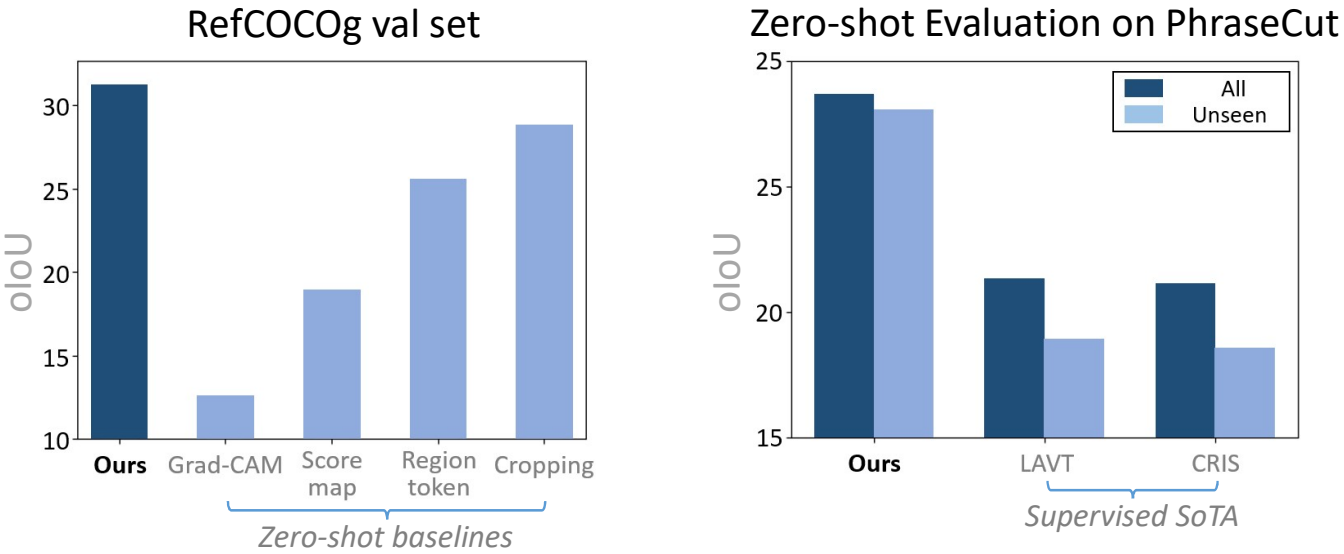


Mask Prediction

Experiments

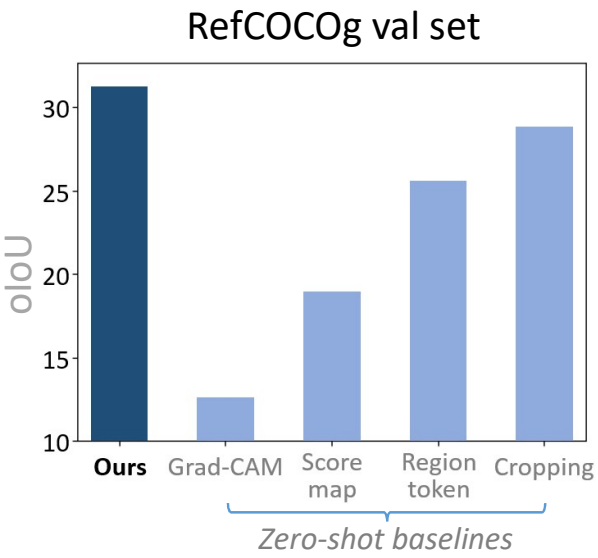


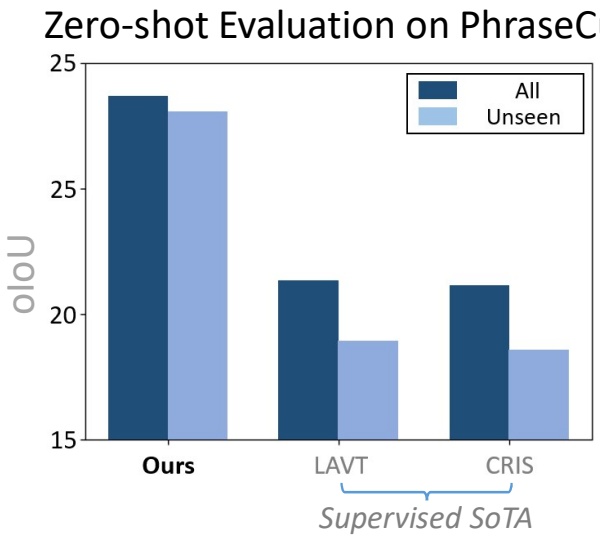
Experiments

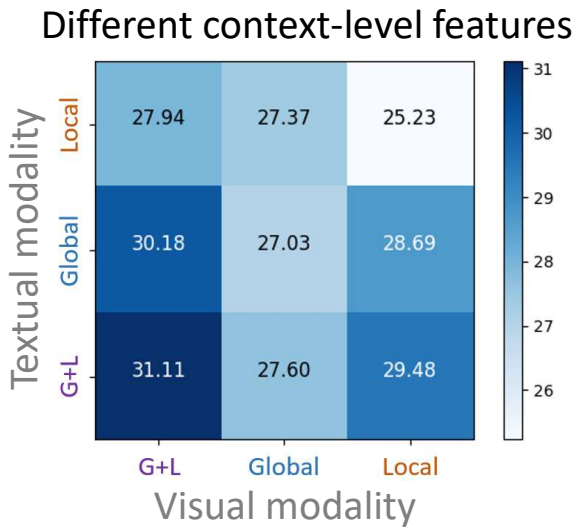


LAVT: Language-aware vision transformer for referring image segmentation. In CVPR, 2022
CRIS: Clip-driven referring image segmentation. In CVPR, 2022

Experiments







LAVT: Language-aware vision transformer for referring image segmentation. In CVPR, 2022
CRIS: Clip-driven referring image segmentation. In CVPR, 2022

Qualitative Results

Different context-levels of visual feature

Qualitative Results

Different context-levels of visual feature

Image

Expression:

```
a green bicycle ridden by a man
in a black windbreaker
```

GT

Local visual

Qualitative Results

Different context-levels of visual feature

Image

Expression:

```
a green bicycle ridden by a man
in a black windbreaker
```

GT

Local visual

Global visual

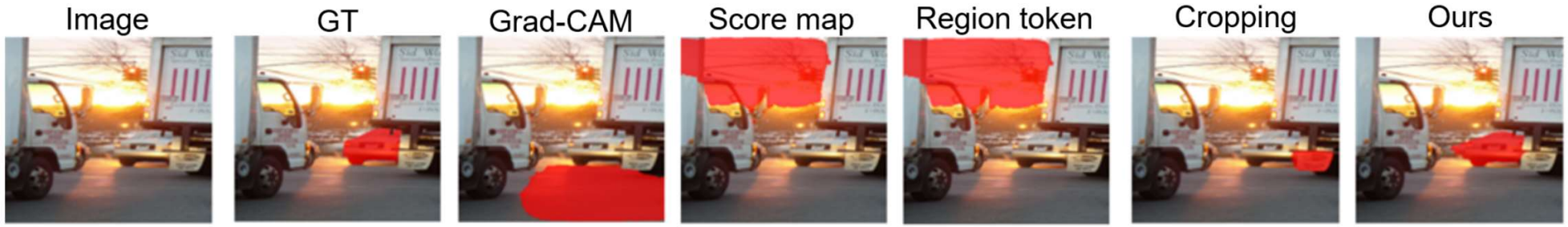
Qualitative Results

Different context-levels of visual feature

Qualitative Results

Different context-levels of textual feature

Qualitative Results with baselines



- Mask generator: FreeSOLO in all baselines

FreeSOLO: Learning to segment without annotations. In CVPR ,2022.

Thanks!
Visit our poster
THU-AM-283