CVPR
JUNE 17-21, 2024
SEATTLE, WA

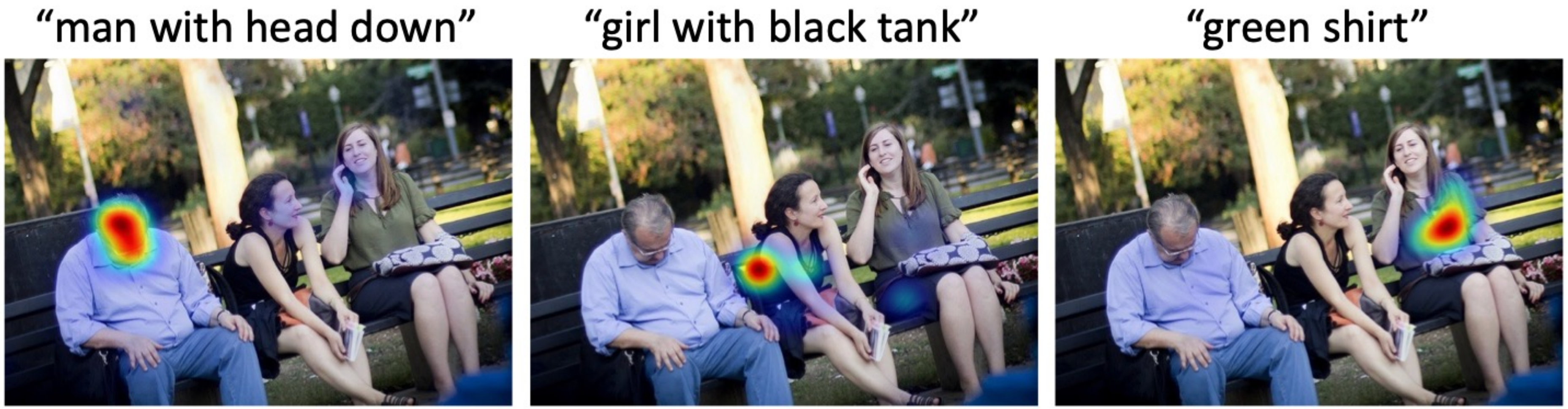
Improved Visual Grounding through Self-Consistent Explanations

Ruozhen He[1], Paola Cascante-Bonilla[1], Ziyan Yang[1], Alexander C. Berg[2], Vicente Ordonez[1]

[1] Rice University [2] University of California, Irvine

Introduction

- Vision-and-Language Models (VLM)
- Visual Grounding through Visual Explanations



“man with head down”

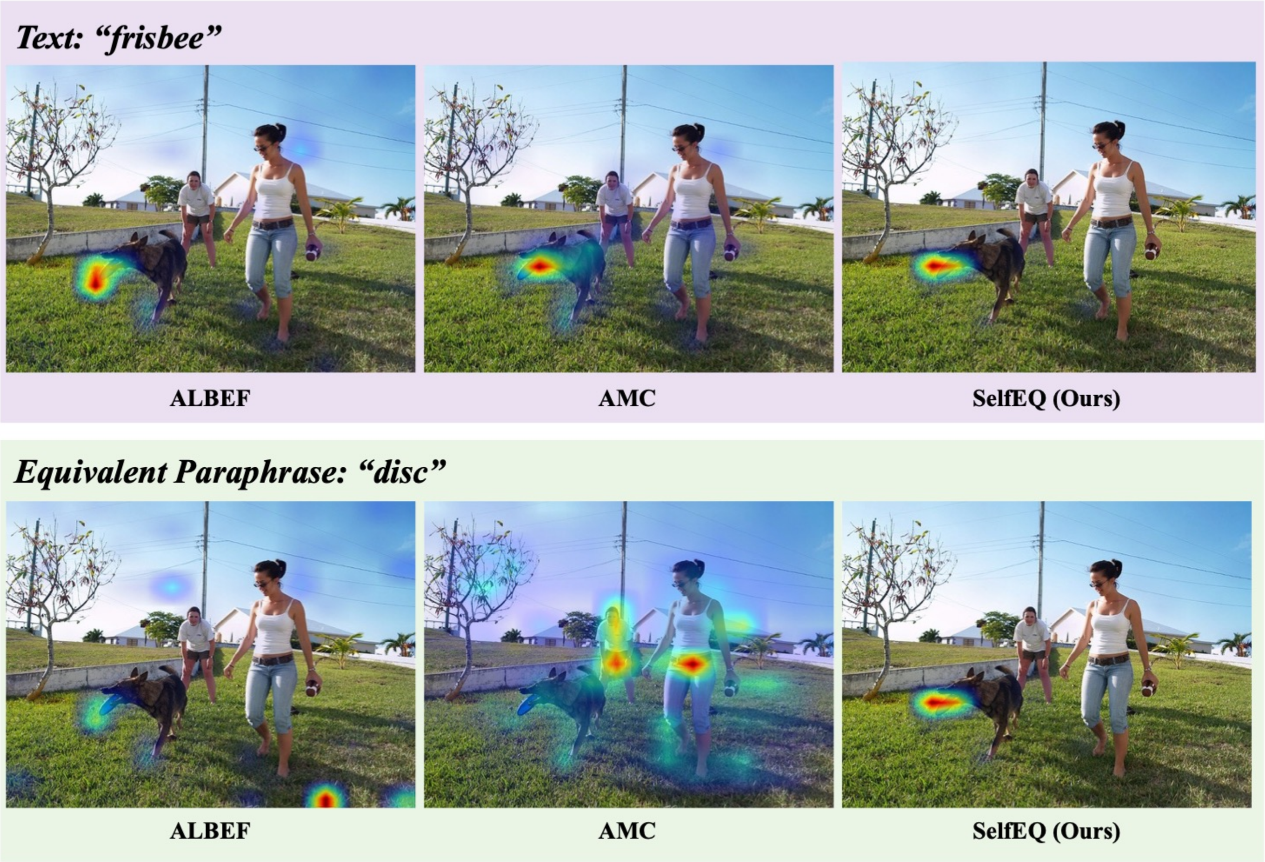
“girl with black tank”

“green shirt”

GradCAM visualization of the ALBEF model.[1]

[1] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, *34*, 9694-9705.

Motivation



- Weakly-Supervised Visual Grounding
  - Without any forms of region annotations

- Higher Self-Consistency
- Better localization
- Larger working vocabulary

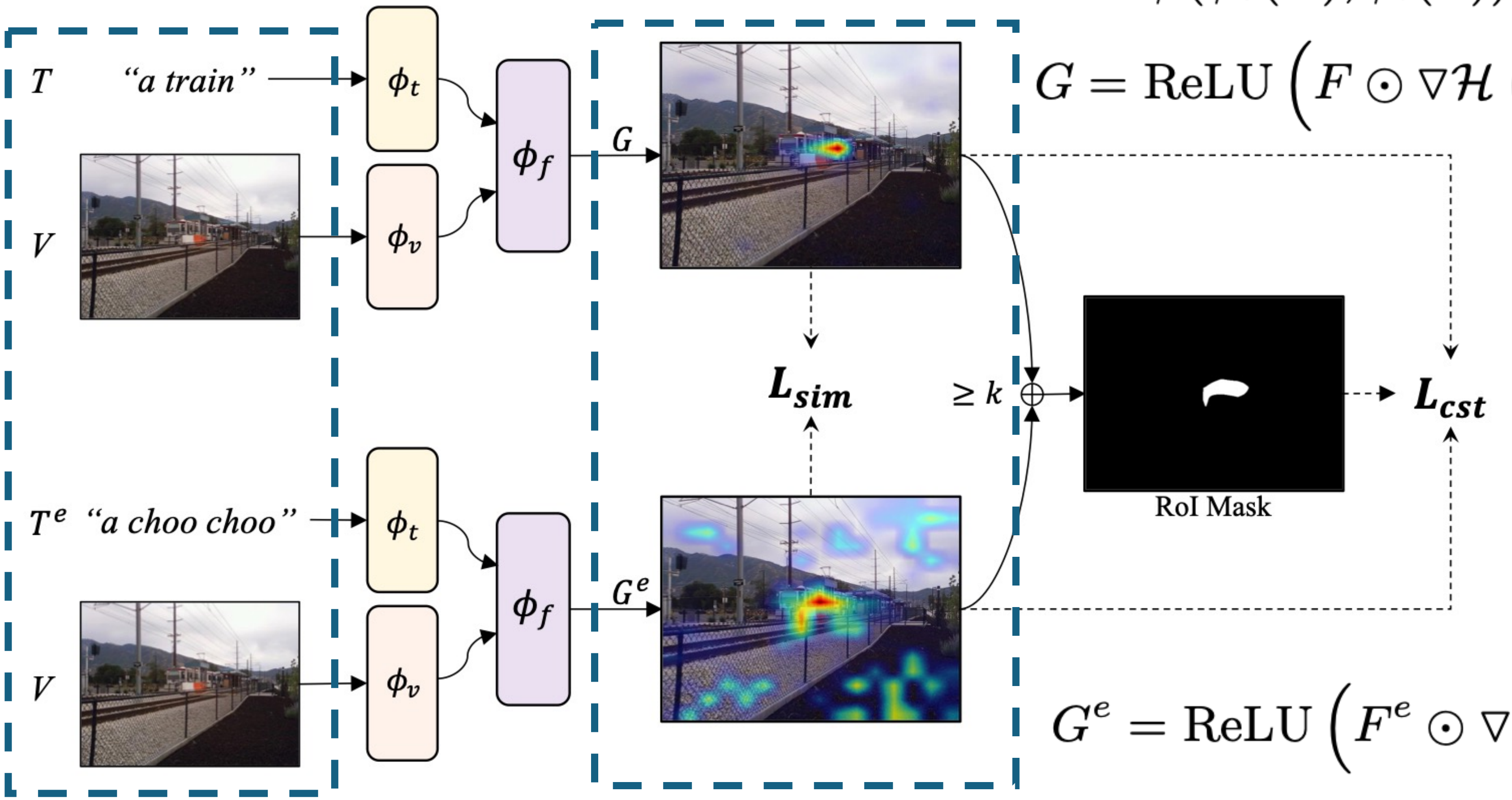
Overview

- Paraphrase Generation
  - Utilize a Large Language Model (LLM) to generate paraphrases.
- **Self**-consistency **EQ**uivalence Tuning (SelfEQ)
  - Weakly-supervised objective.
  - Encourages consistent visual explanations.
  - Applies to paraphrased input text pairs that refer to the same object or region in an image.

Self-Consistent **EQ**uivalent Tuning Objective

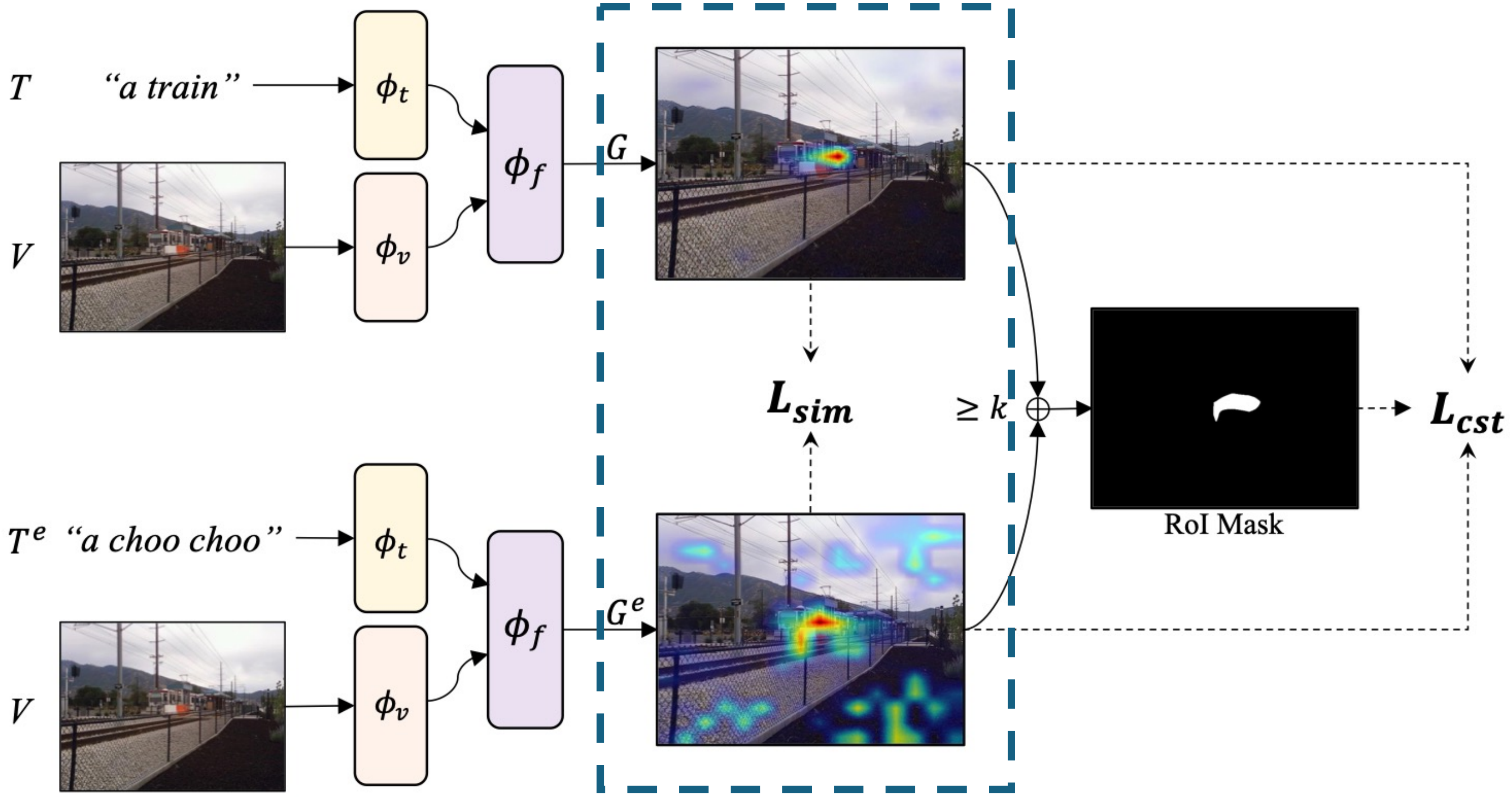
$$F = \phi\left(\phi_v(V), \phi_t(T)\right), F^e = \phi\left(\phi_v(V), \phi_t(T^e)\right)$$

$$G = \mathrm{ReLU}\left(F \odot \nabla\mathcal{H}\left(\vec{y}, \phi_f^{cls}\left(\phi_v(V), \phi_t(T)\right)\right)\right)$$



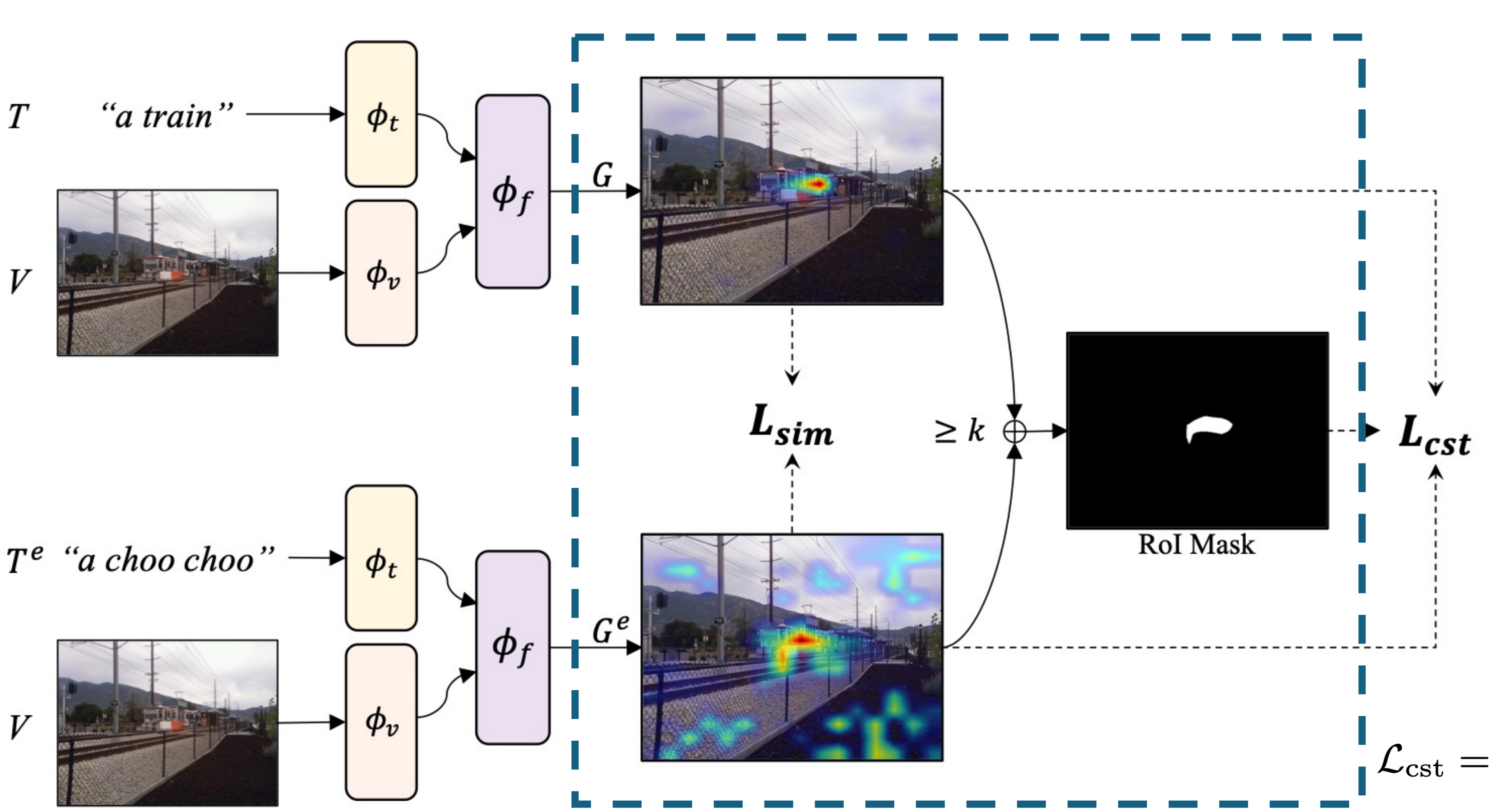
$$G^e = \mathrm{ReLU}\left(F^e \odot \nabla\mathcal{H}\left(\vec{y}, \phi_f^{cls}\left(\phi_v(V), \phi_t(T^e)\right)\right)\right)$$

Self-Consistent **EQ**uivalent Tuning Objective



$$\mathcal{L}_{\text{sim}} = \mathbb{E}_{(V,T,T^e)\sim D'}\left[\frac{1}{N}\sum_{i,j}(G_{i,j} - G^e_{i,j})^2\right]$$

Self-Consistent **EQ**uivalent Tuning Objective



$$M_{i,j} = \begin{cases} 1, (G_{i,j} + G^e_{i,j}) \geq k \\ 0, (G_{i,j} + G^e_{i,j}) < k \end{cases}$$

$$R = G \odot M, \quad R^e = G^e \odot M.$$

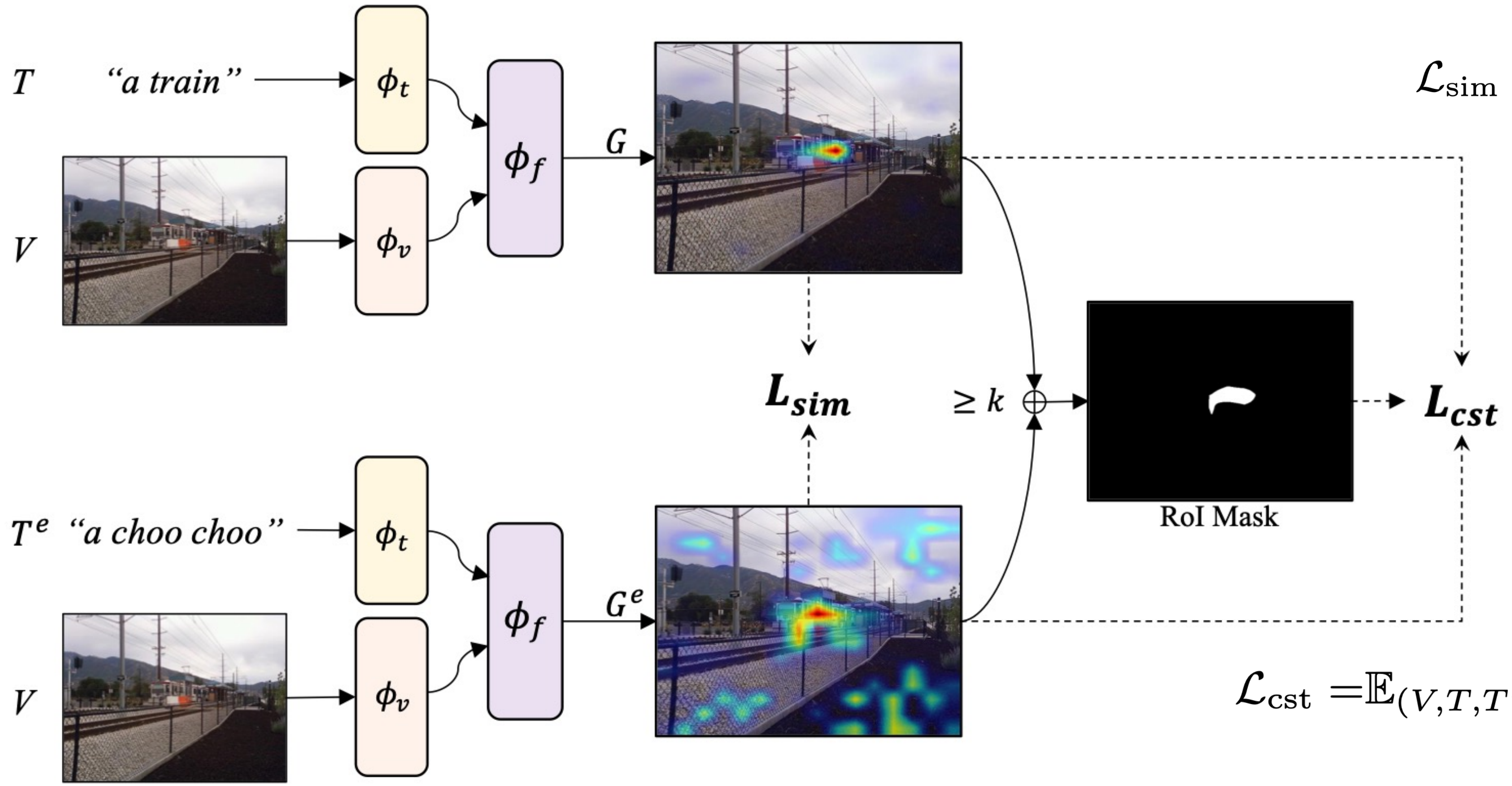
$$\mu_{RoI} = \frac{\sum_{i,j} R_{i,j}}{\sum_{i,j} M_{i,j}}, \mu^e_{RoI} = \frac{\sum_{i,j} R^e_{i,j}}{\sum_{i,j} M_{i,j}}$$

$$\sigma_{RoI} = \sqrt{\frac{\sum_{i,j} M_{i,j} \cdot (R_{i,j} - \mu_{RoI})^2}{\sum_{i,j} M_{i,j}}}$$

$$\sigma^e_{RoI} = \sqrt{\frac{\sum_{i,j} M_{i,j} \cdot (R^e_{i,j} - \mu^e_{RoI})^2}{\sum_{i,j} M_{i,j}}}$$

$$\mathcal{L}_{\text{cst}} = \mathbb{E}_{(V,T,T^e)\sim D'} \Big[\sigma_{RoI} + \sigma^e_{RoI} + \max(0, k/2 - \mu_{RoI}) + \max(0, k/2 - \mu^e_{RoI}) \Big]$$

Self-Consistent **EQ**uivalent Tuning Objective

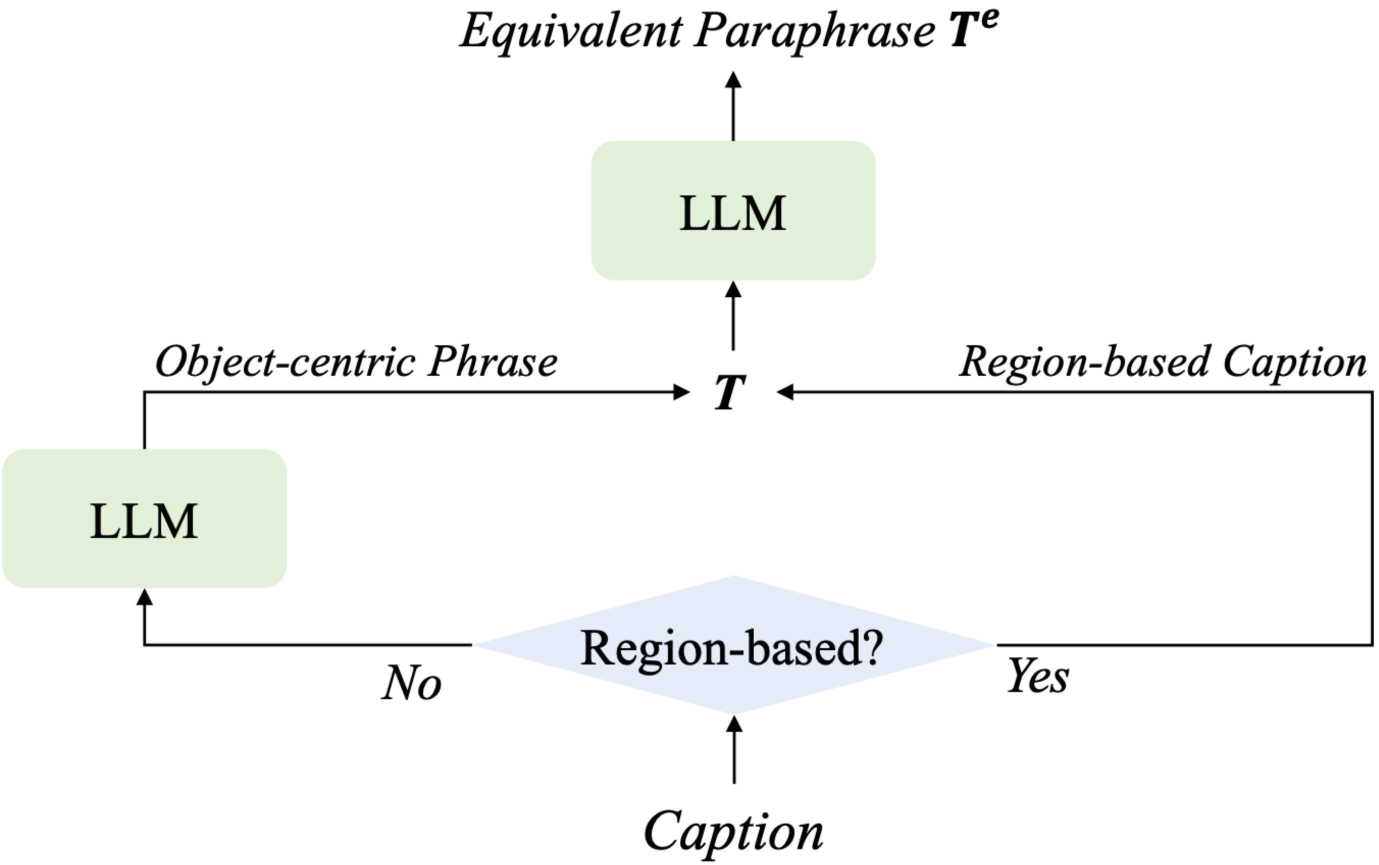


$$\mathcal{L}_{\text{sim}} = \mathbb{E}_{(V,T,T^e)\sim D'}\left[\frac{1}{N}\sum_{i,j}(G_{i,j} - G^e_{i,j})^2\right]$$

$$\mathcal{L}_{\text{cst}} = \mathbb{E}_{(V,T,T^e)\sim D'}\Big[\sigma_{RoI} + \sigma^e_{RoI} + \max(0, k/2 - \mu_{RoI}) + \max(0, k/2 - \mu^e_{RoI})\Big]$$

$$\mathcal{L}_{\text{SelfEQ}} = \mathcal{L}_{\text{sim}} + \lambda \cdot \mathcal{L}_{\text{cst}}$$

Self-Consistency Data Augmentation



- Phrase Extraction

  Break down global captions into object-centric phrases.

- Paraphrase Strategy

  Replace primary objects while keeping other attributes intact.

Data Examples

Q: "a stadium"

```
A: { "group": "stadium",
     "synonym": ["arena", "sports stadium"],
     "antonym": [],
     "hypernym": ["sports facility", "outdoor sports venue", "place of sporting events"],
     "meronym": ["stadium seating", "stadium lights", "stadium sound system"]}
```

Region-based Caption

Q: "A white and gray cat laying underneath an umbrella."

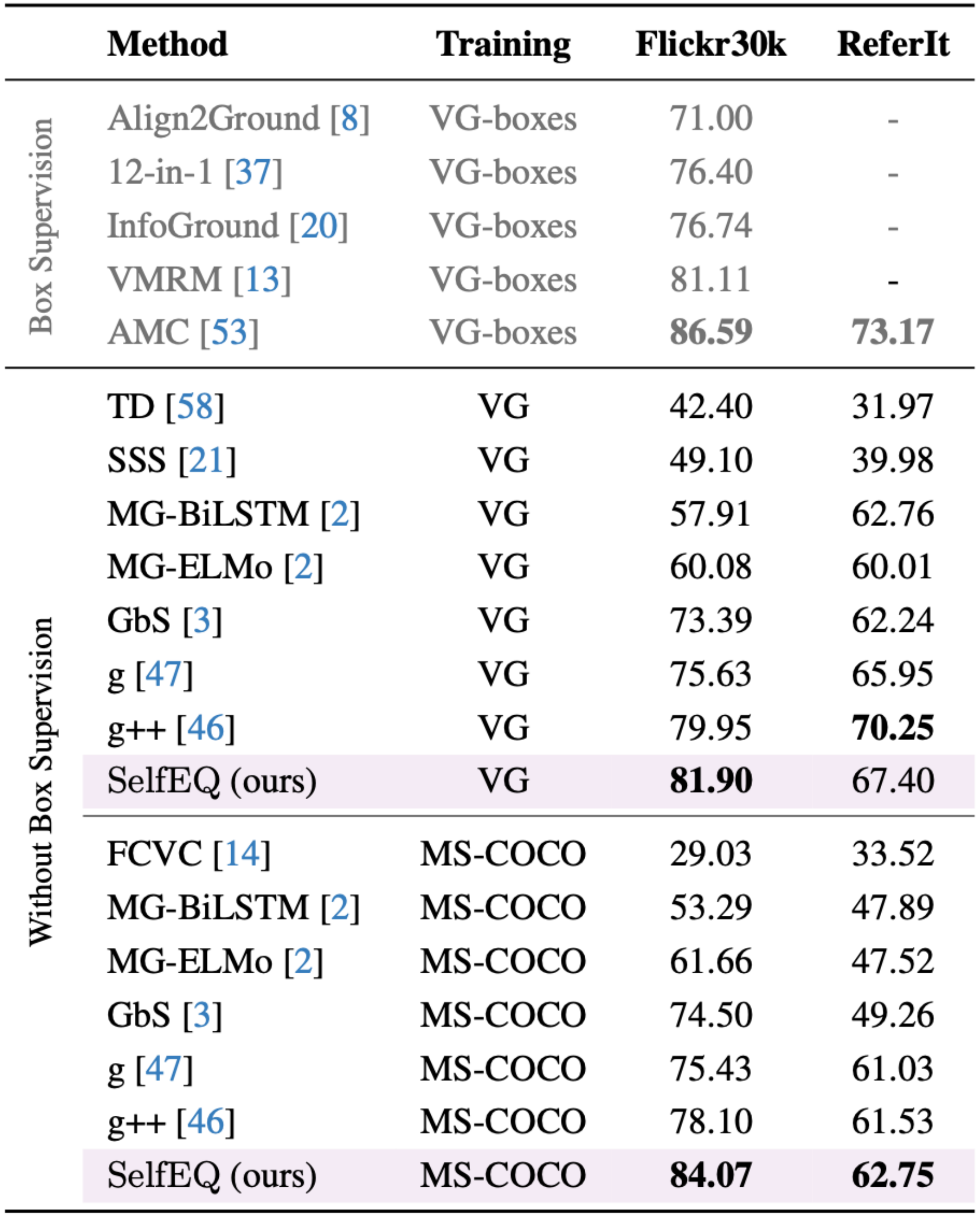
```
A1: "object-centric phrases": ["a white and gray cat", "an umbrella"]
A2: {"group": "cat",
     "synonym": ["feline", "cat animal", "cat creature"],
     "antonym": [],
     "hypernym": ["mammal", "animal", "pet"],
     "meronym": ["cat fur", "cat eyes", "cat nose", "cat paws", "cat tail", "cat body",
                 "cat head", "cat ears"]}
    {"group": "umbrella",
     "synonym": ["parasol", "brolly"],
     "antonym": ["sun"],
     "hypernym": ["covering", "shelter", "protection", "canopy"],
     "meronym": ["umbrella handle", "umbrella frame", "umbrella fabric",
                 "umbrella spike"]}
```

Global-based Caption

Experiments

- Training
  - Visual Genome (VG)
  - MS-COCO
- Evaluation
  - Flickr30k
  - ReferIt
- Metric
  - Pointing Game Accuracy

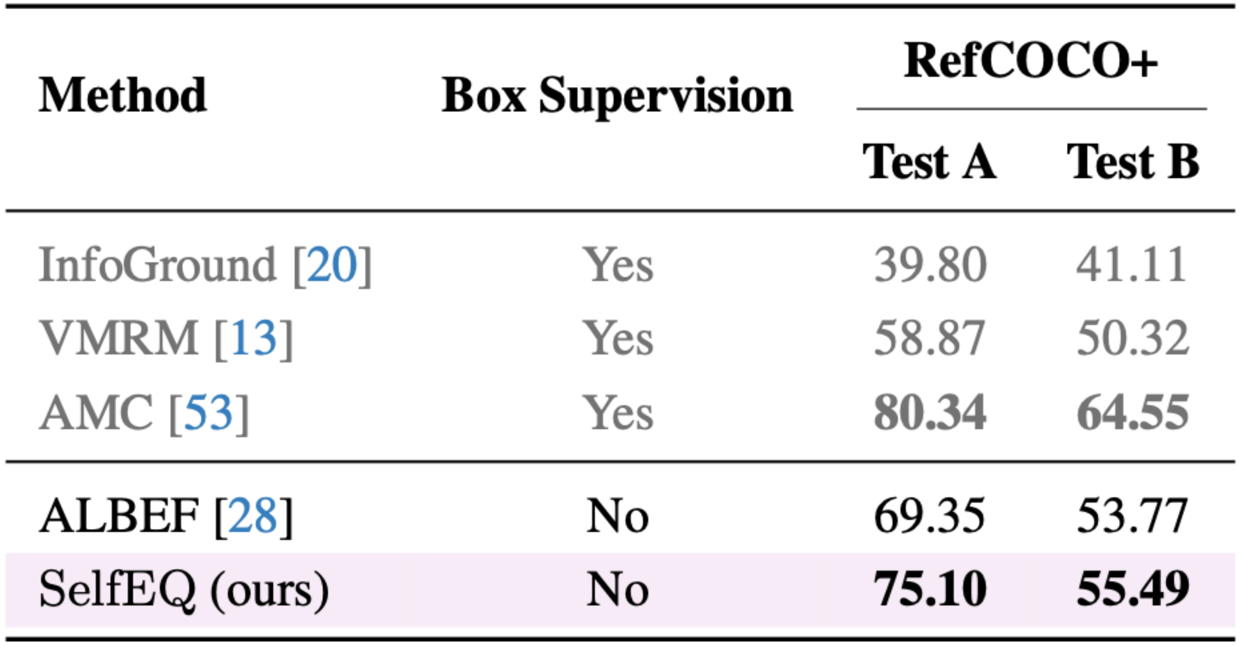
Experiments



	Method	Training	Flickr30k	ReferIt
Box Supervision	Align2Ground [8]	VG-boxes	71.00	-
	12-in-1 [37]	VG-boxes	76.40	-
	InfoGround [20]	VG-boxes	76.74	-
	VMRM [13]	VG-boxes	81.11	-
	AMC [53]	VG-boxes	**86.59**	**73.17**
Without Box Supervision	TD [58]	VG	42.40	31.97
	SSS [21]	VG	49.10	39.98
	MG-BiLSTM [2]	VG	57.91	62.76
	MG-ELMo [2]	VG	60.08	60.01
	GbS [3]	VG	73.39	62.24
	g [47]	VG	75.63	65.95
	g++ [46]	VG	79.95	**70.25**
	SelfEQ (ours)	VG	**81.90**	67.40
	FCVC [14]	MS-COCO	29.03	33.52
	MG-BiLSTM [2]	MS-COCO	53.29	47.89
	MG-ELMo [2]	MS-COCO	61.66	47.52
	GbS [3]	MS-COCO	74.50	49.26
	g [47]	MS-COCO	75.43	61.03
	g++ [46]	MS-COCO	78.10	61.53
	SelfEQ (ours)	MS-COCO	**84.07**	**62.75**

Table 1. Visual Grounding results on two benchmarks using pointing game accuracy with two training datasets.

Experiments



Method	Box Supervision	RefCOCO+	
		Test A	Test B
InfoGround [20]	Yes	39.80	41.11
VMRM [13]	Yes	58.87	50.32
AMC [53]	Yes	**80.34**	**64.55**
ALBEF [28]	No	69.35	53.77
SelfEQ (ours)	No	**75.10**	**55.49**

Table 2. Results on RefCOCO+ pointing game accuracy.

Ablation Studies: Data Quantity

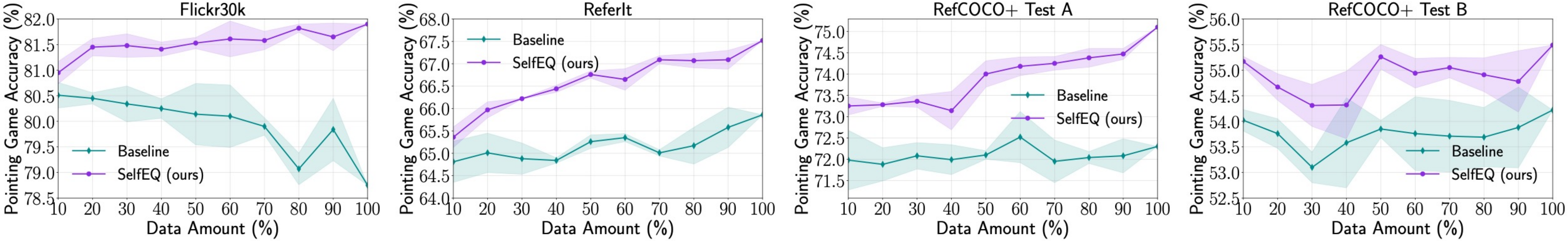
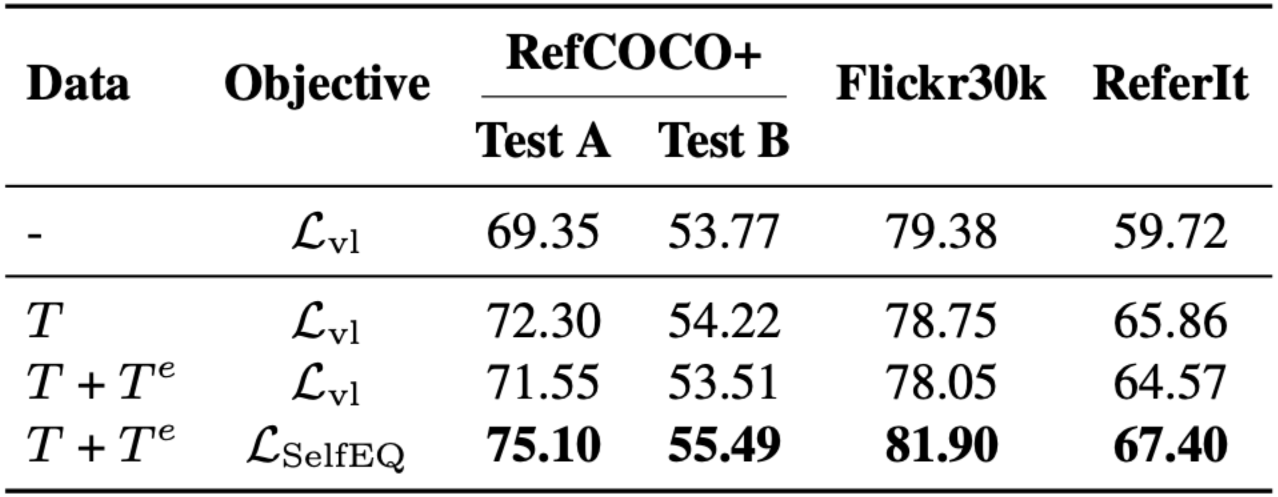


Figure 6. Tuning performance with different data quantities on Flickr30k, ReferIt, RefCOCO+ Test A and Test B.

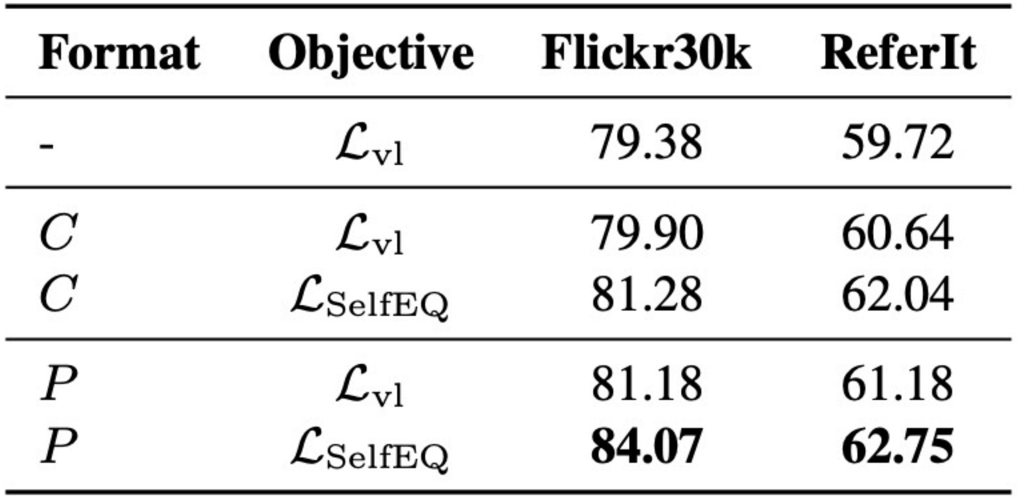
Ablation Studies: Data Quantity



Data	Objective	RefCOCO+		Flickr30k	ReferIt
		Test A	Test B		
-	$\mathcal{L}_{\text{vl}}$	69.35	53.77	79.38	59.72
T	$\mathcal{L}_{\text{vl}}$	72.30	54.22	78.75	65.86
$T + T^e$	$\mathcal{L}_{\text{vl}}$	71.55	53.51	78.05	64.57
$T + T^e$	$\mathcal{L}_{\text{SelfEQ}}$	**75.10**	**55.49**	**81.90**	**67.40**

Table 3. Ablation studies on different ways to utilize extra equivalent paraphrased data.

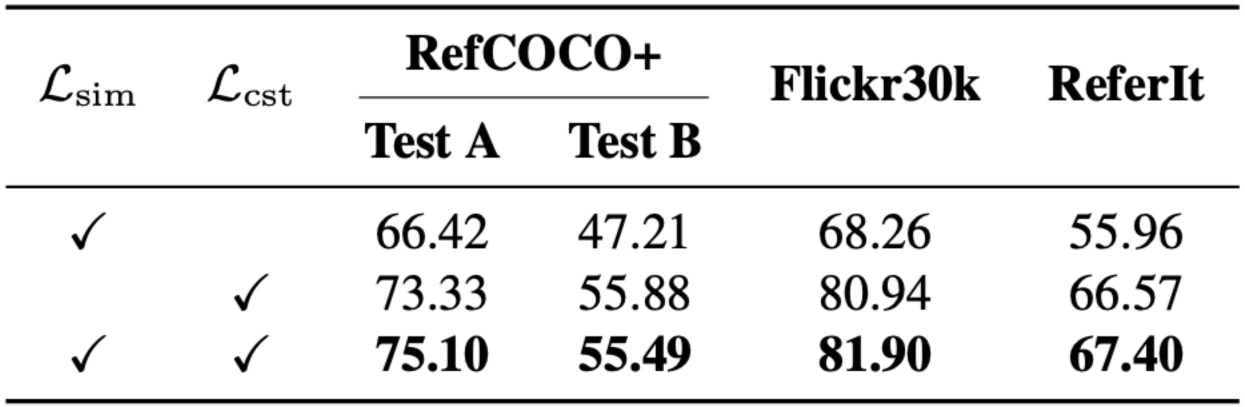
Ablation Studies: Data Augmentation



Format	Objective	Flickr30k	ReferIt
-	$\mathcal{L}_{\mathrm{vl}}$	79.38	59.72
C	$\mathcal{L}_{\mathrm{vl}}$	79.90	60.64
C	$\mathcal{L}_{\mathrm{SelfEQ}}$	81.28	62.04
P	$\mathcal{L}_{\mathrm{vl}}$	81.18	61.18
P	$\mathcal{L}_{\mathrm{SelfEQ}}$	**84.07**	**62.75**

Table 4. Comparisons on data augmentation strategies or global based captions in MS-COCO.

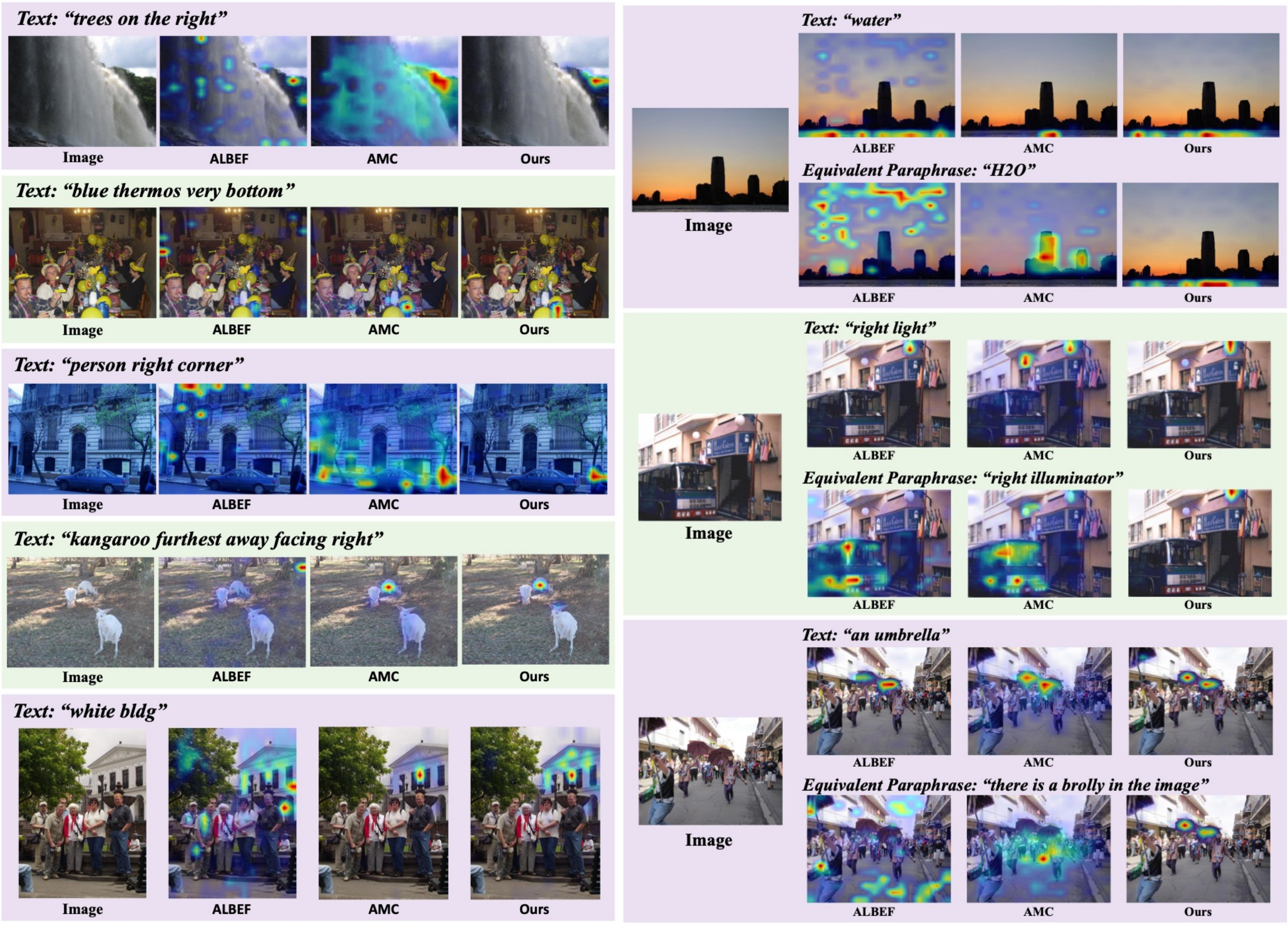
Ablation Studies: Objective



$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{cst}}$	**RefCOCO+**		**Flickr30k**	**ReferIt**
		Test A	**Test B**		
✓		66.42	47.21	68.26	55.96
	✓	73.33	55.88	80.94	66.57
✓	✓	**75.10**	**55.49**	**81.90**	**67.40**

Table 5. Ablation studies on objective component of self-consistency equivalence tuning objective L_{SelfEQ}.

Qualitative Results


Text: "trees on the right"
Image
ALBEF
AMC
Ours
Text: "blue thermos very bottom"
Image
ALBEF
AMC
Ours
Text: "person right corner"
Image
ALBEF
AMC
Ours
Text: "kangaroo furthest away facing right"
Image
ALBEF
AMC
Ours
Text: "white bldg"
Image
ALBEF
AMC
Ours
Text: "water"
Image
ALBEF
AMC
Ours
Equivalent Paraphrase: "H2O"
ALBEF
AMC
Ours
Text: "right light"
Image
ALBEF
AMC
Ours
Equivalent Paraphrase: "right illuminator"
ALBEF
AMC
Ours
Text: "an umbrella"
Image
ALBEF
AMC
Ours
Equivalent Paraphrase: "there is a brolly in the image"
ALBEF
AMC
Ours

RICE

UCIrvine
University of California · Irvine

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Thank You!

Ruozhen He

Paola Cascante-Bonilla

Ziyan Yang

Alexander C. Berg

Vicente Ordonez