RICE

Adobe

Arxiv: https: //arxiv.org/abs/2206.15462
Code: https://github.com/uvavision/AMC-grounding
Demo: https://vislang.ai/amc

Improving Visual Grounding by Encouraging Consistent Gradient-based Explanations

Ziyan Yang[✽]   Kushal Kafle[⌘]   Franck Dernoncourt[⌘]   Vicente Ordonez[✽]

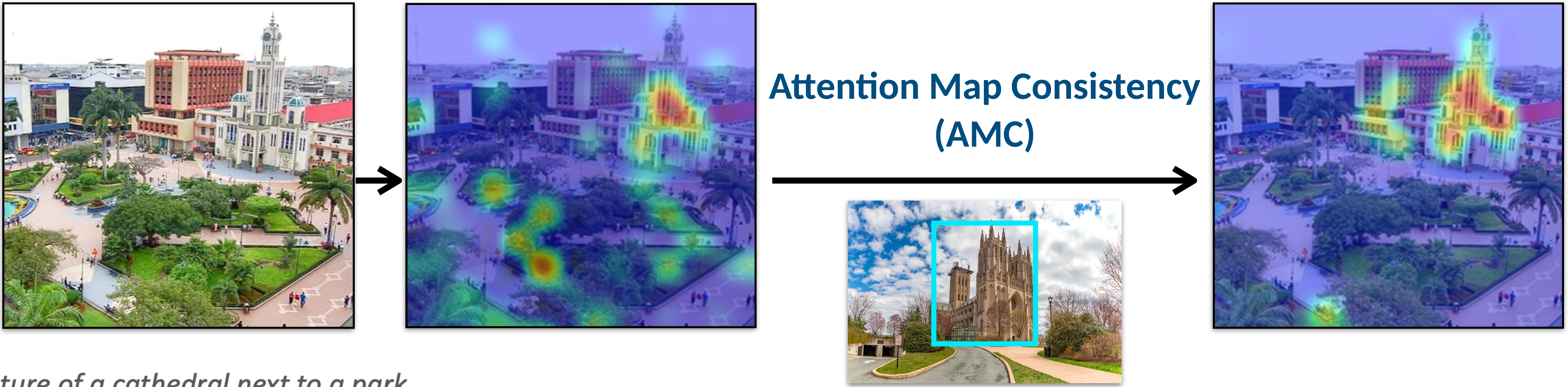
[✽]Rice University
[⌘]Adobe Research

CVPR 2023

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Session & Poster ID: THU-AM-256

Overview



Input Image + Text

Regular V-L Model Explanation

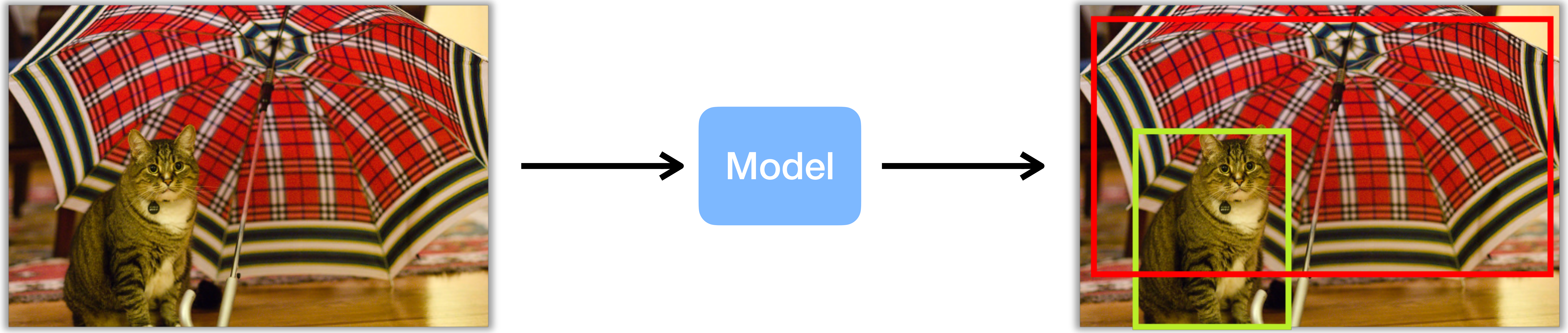
Attention Map Consistency (AMC)

A picture of a cathedral next to a park

Human Explanation

Visual Grounding

- Locate the most relevant region corresponding to a given query

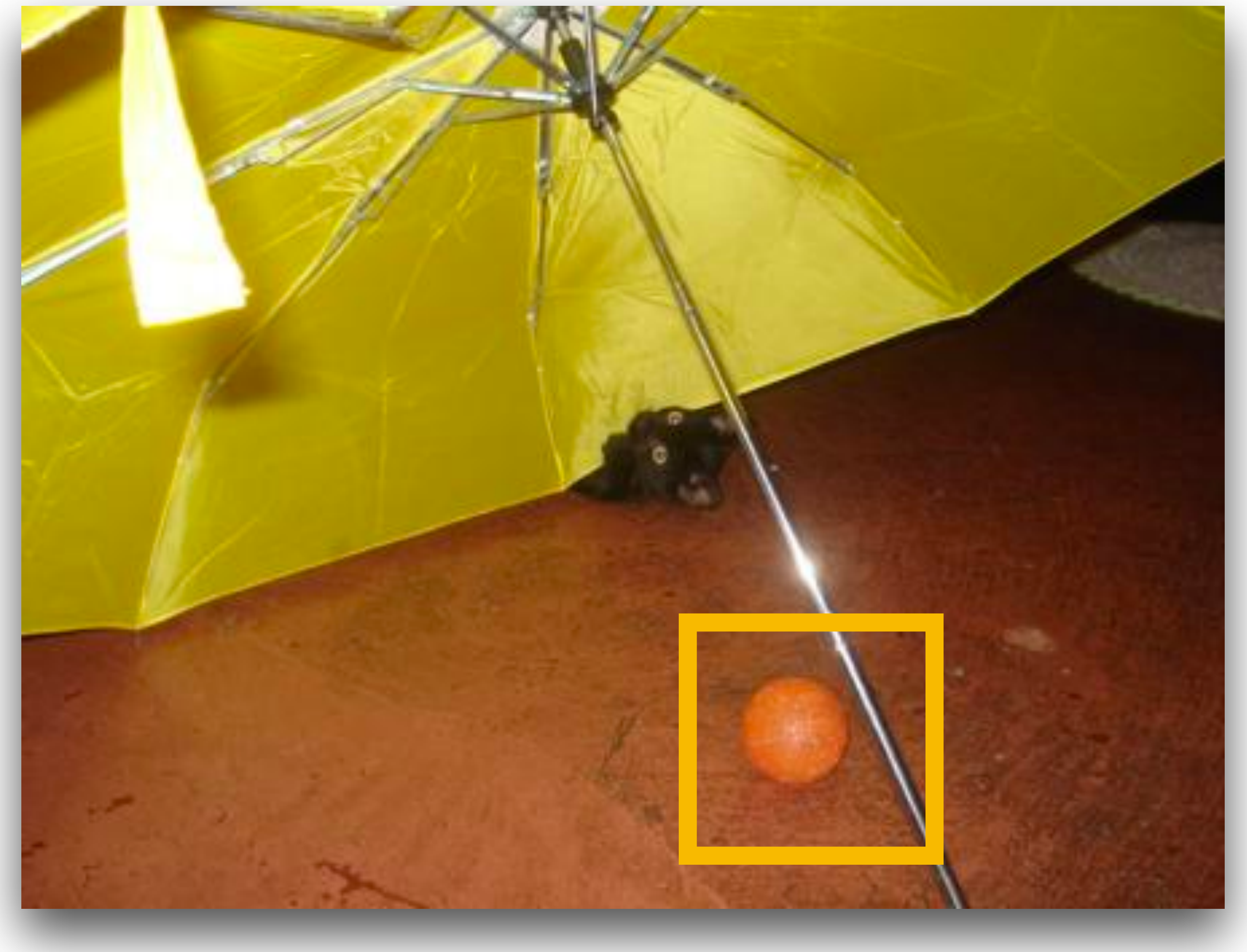


a cat is sitting under a red umbrella

Visual Grounding

a cat under an umbrella

Object Detectors
Annotations: bounding boxes

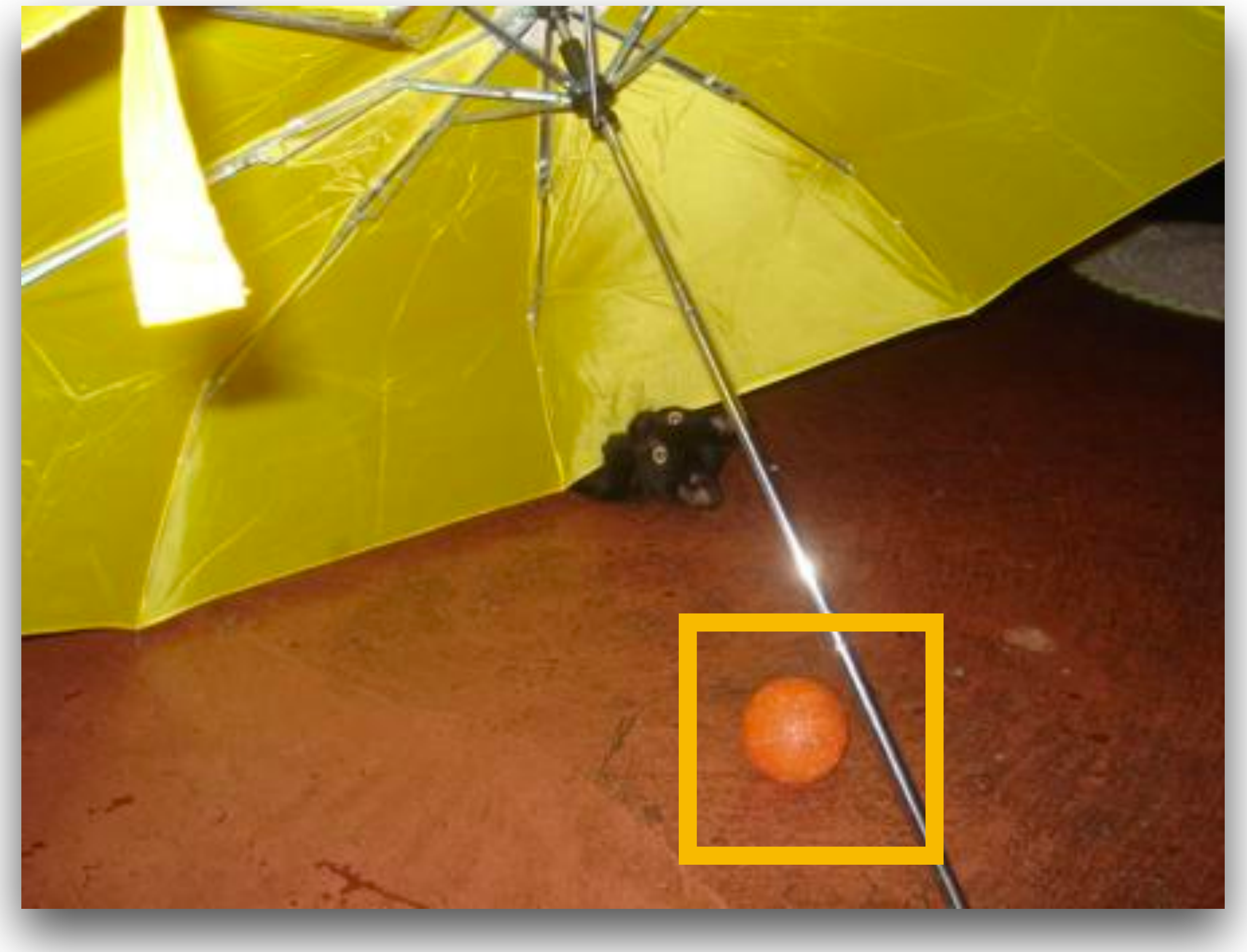


😠

Visual Grounding

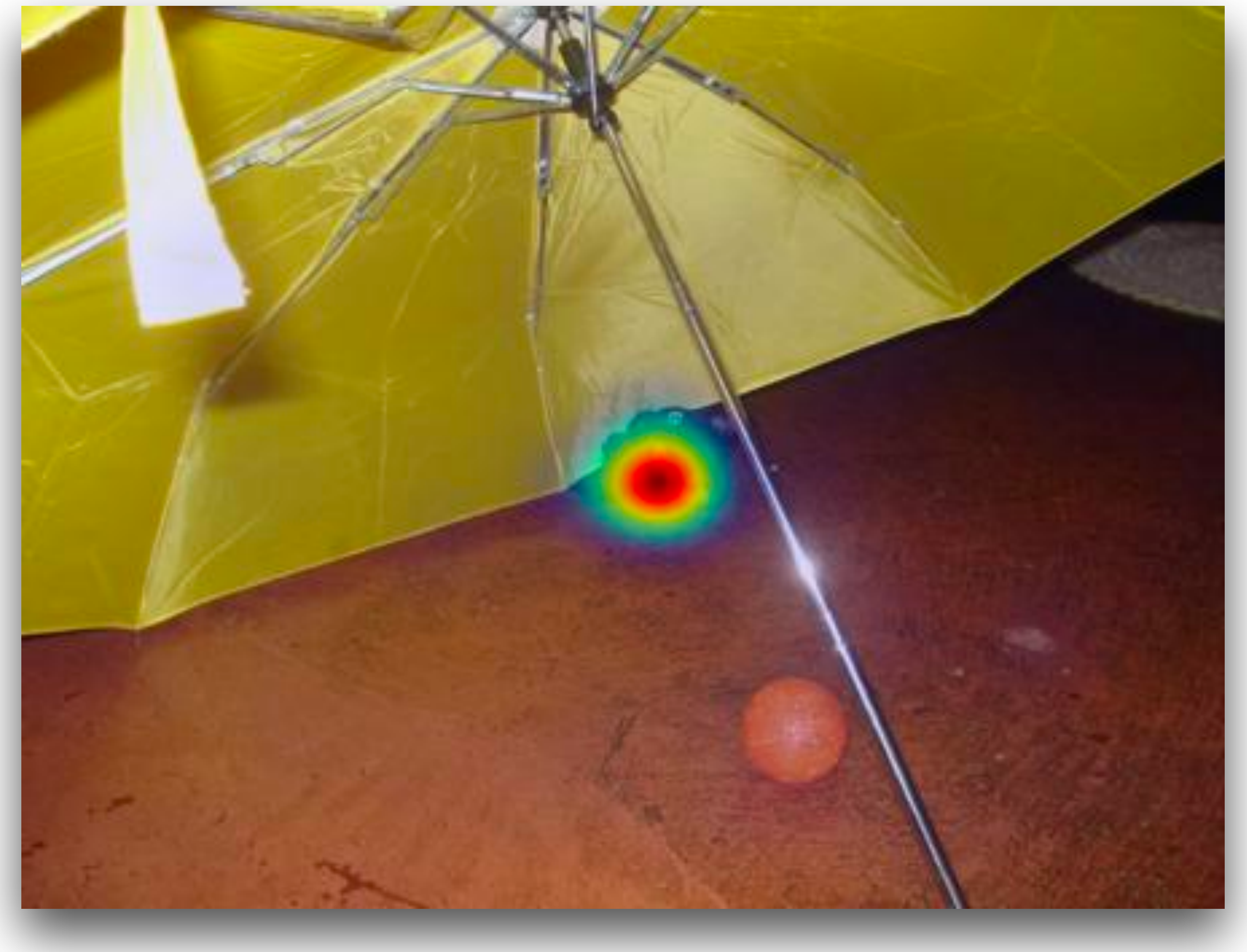
a cat under an umbrella

Object Detectors
Annotations: bounding boxes



😠

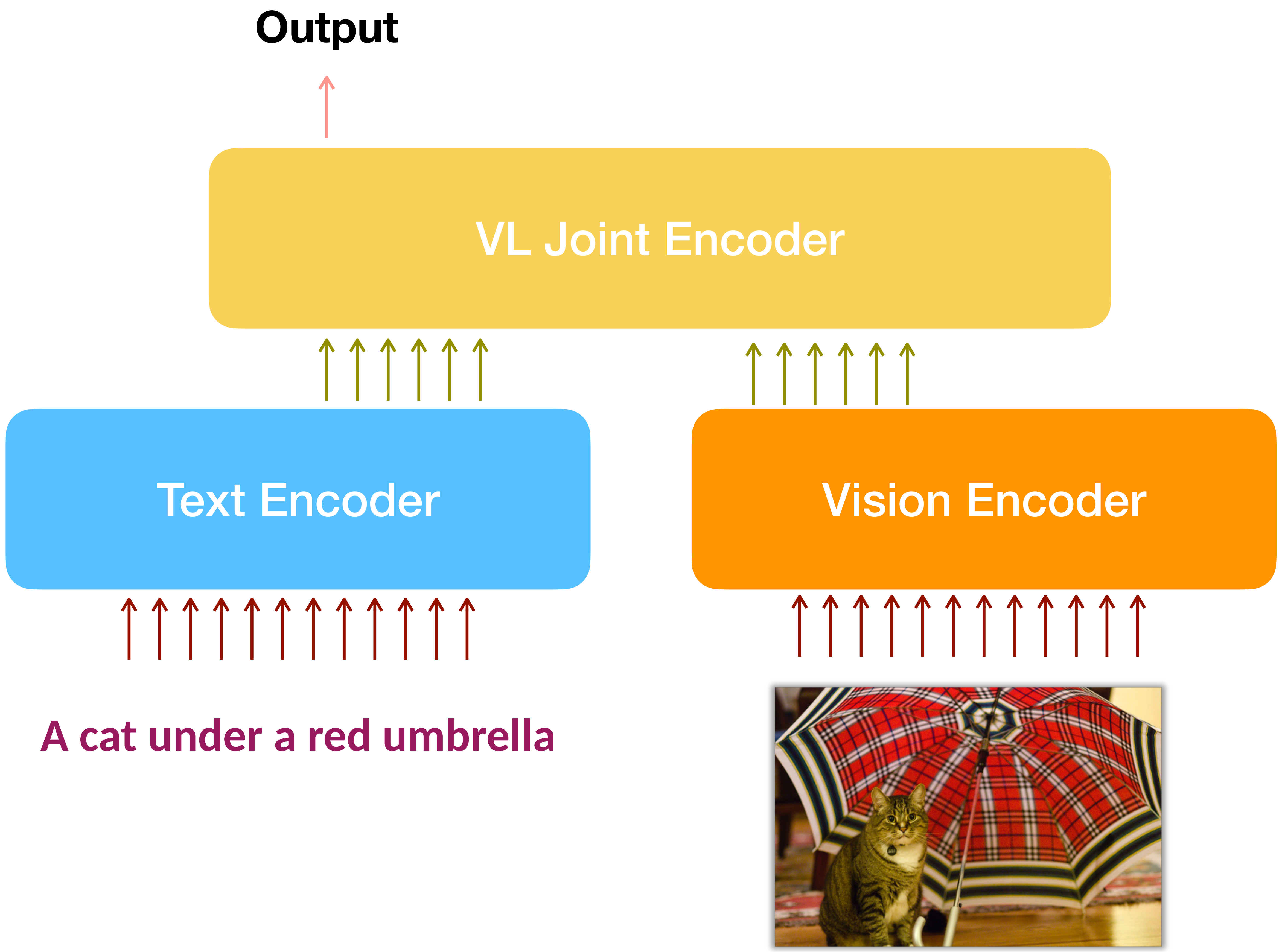
Vision-Language Models
Annotations: bounding boxes



😇

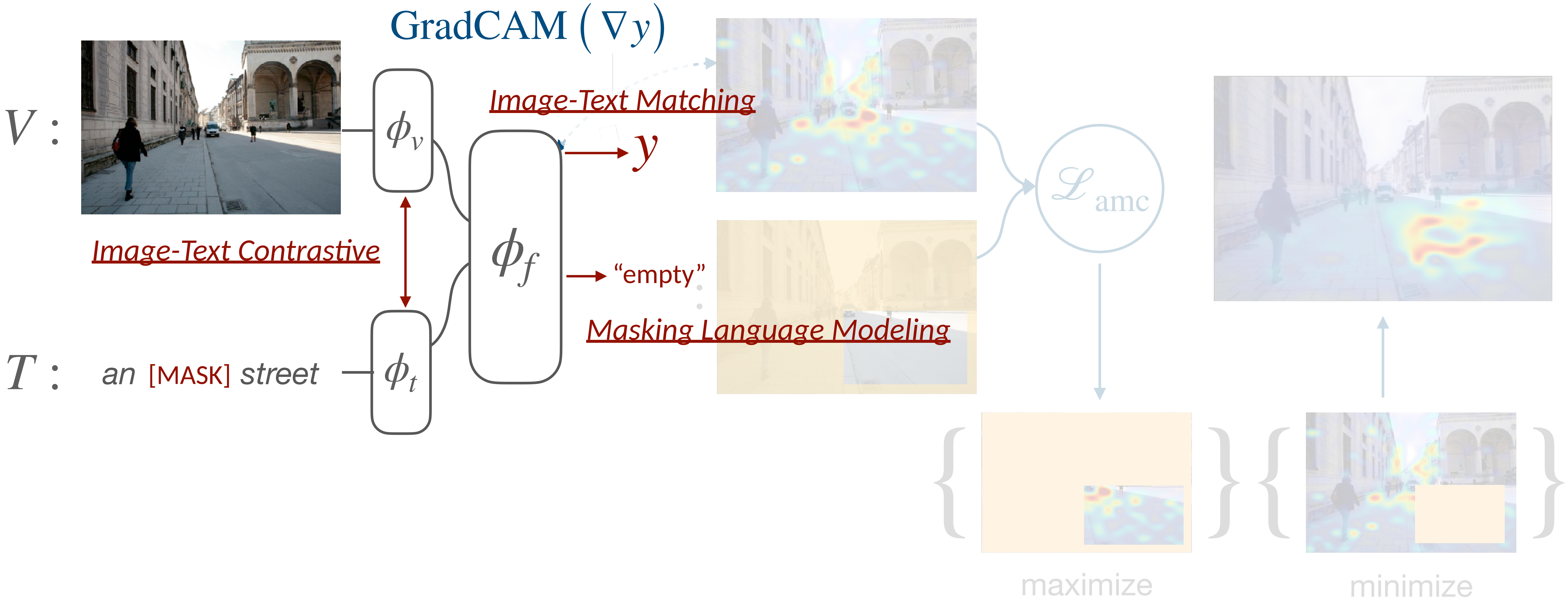
Pre-trained VLMs

- Encoder-Decoder: ALBEF



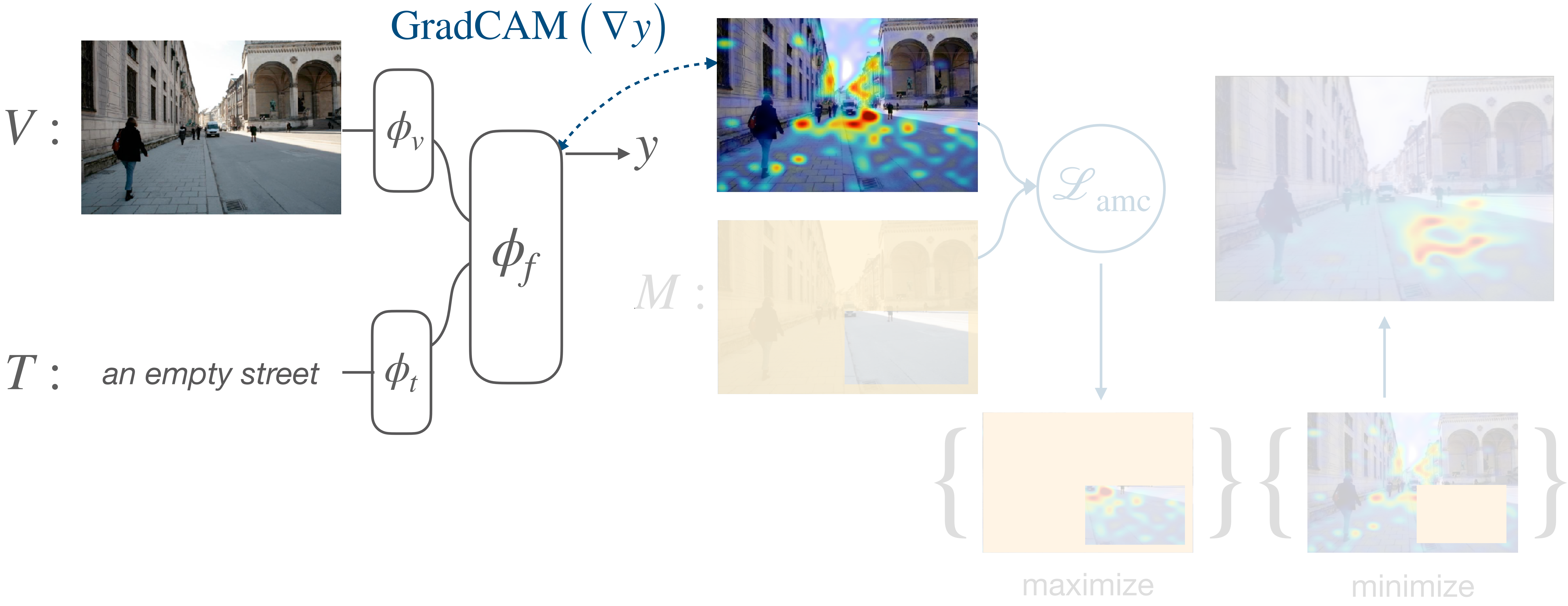
Overview — our method

- Pretraining – from ALBEF
- Assume each sample has: V, T



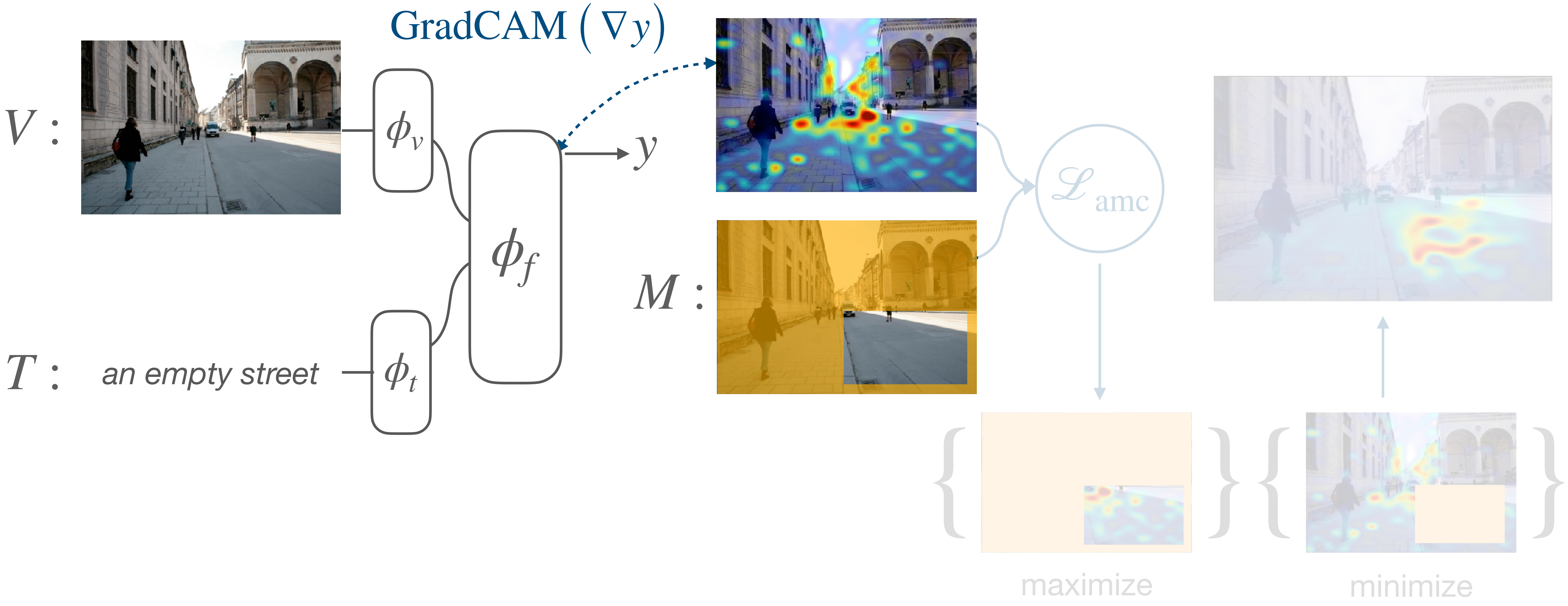
Overview — our method

- Pretraining – from ALBEF
- Assume each sample has: V, T



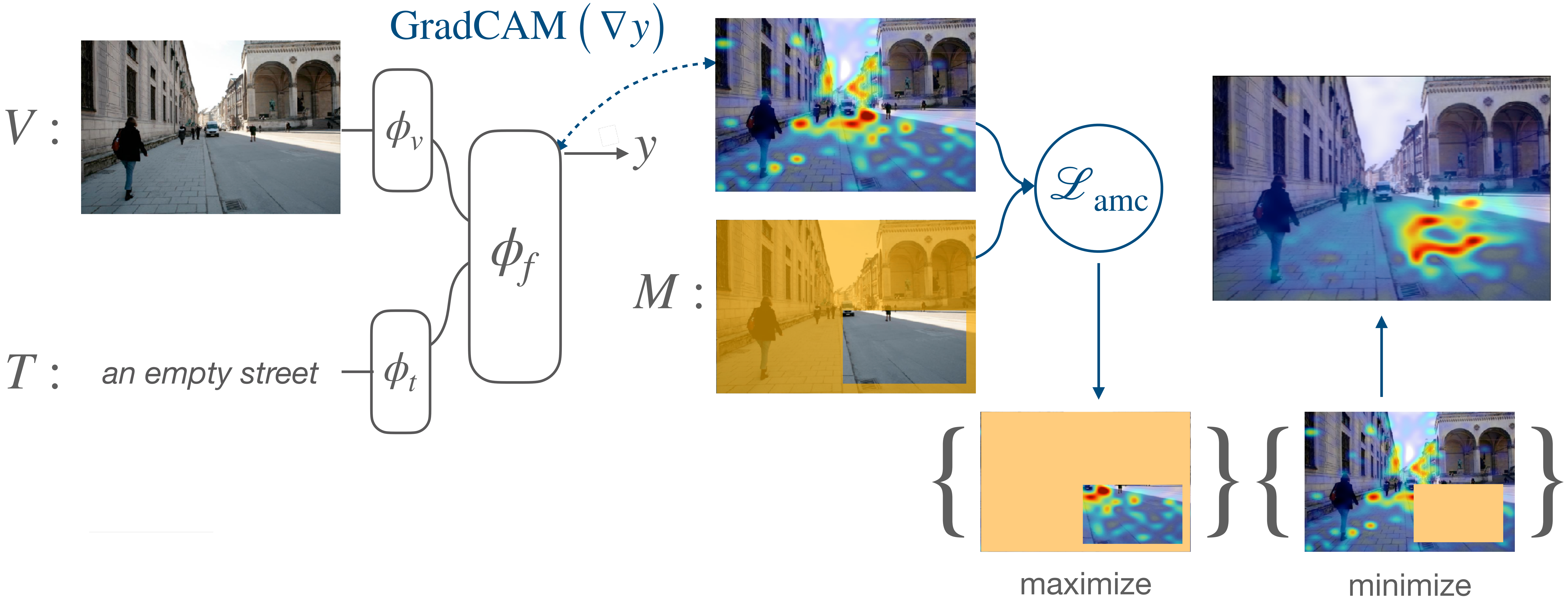
Overview — our method

- Pretraining – from ALBEF
- Assume each sample has: V, T, M



Overview — our method

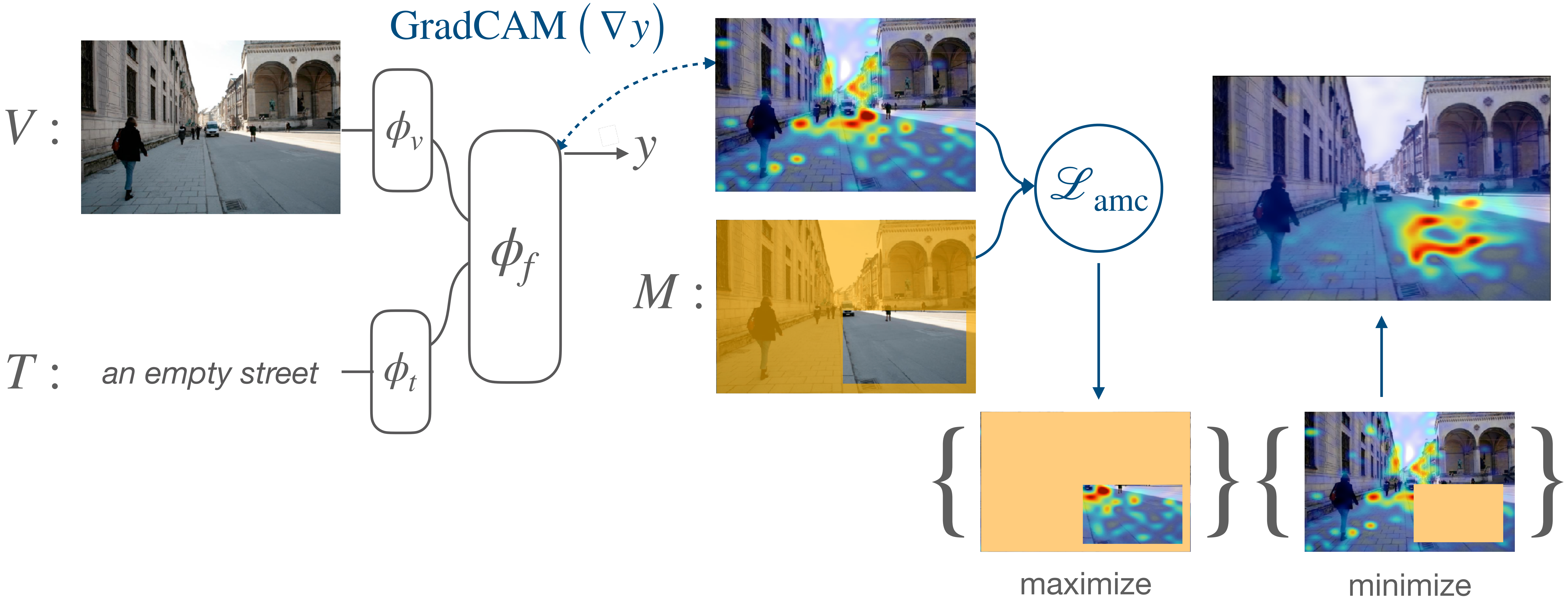
- Pretraining – from ALBEF
- Assume each sample has: V, T, M



Overview — Attention Map Consistency (AMC)

- Pretraining – from ALBEF
- Assume each sample has: V, T, M

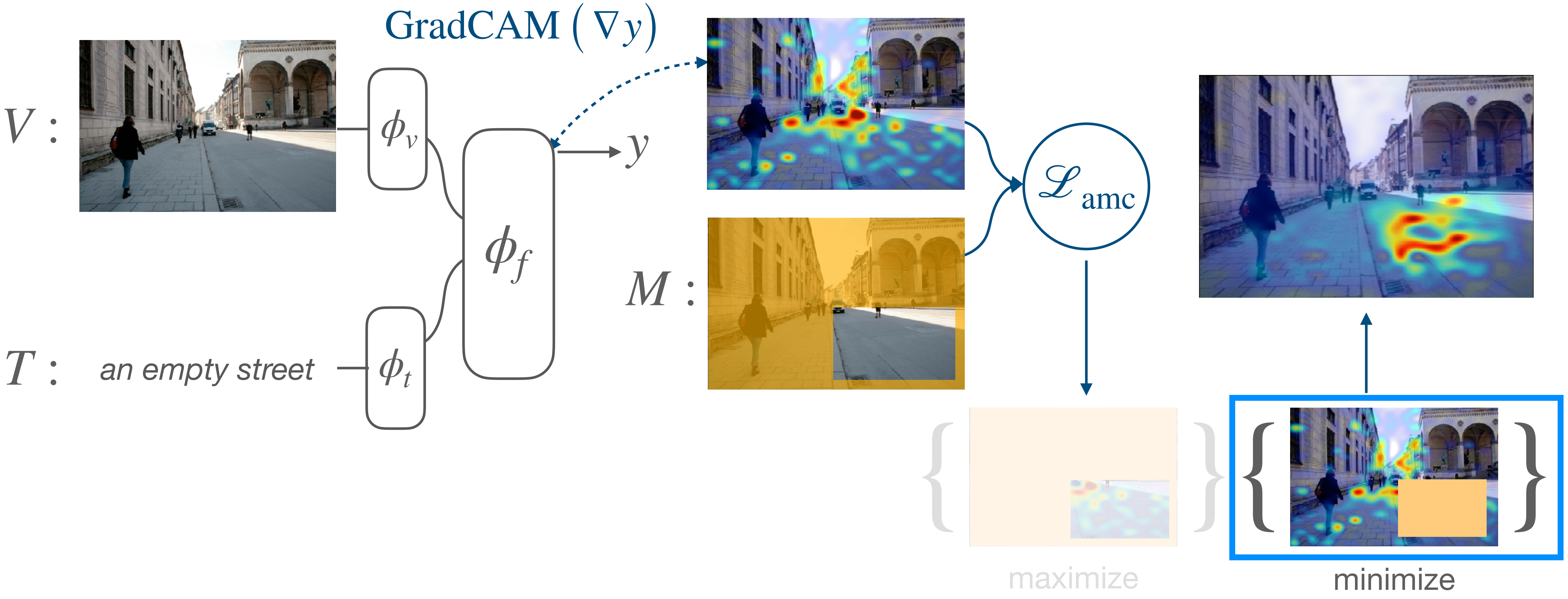
$$\mathcal{L}_{\text{mean}} = \max\left(0, \frac{1}{\sum_{i,j}(1-M_{i,j})}\sum_{i,j}((1-M_{i,j})\,A_{i,j}) - \frac{1}{\sum_{i,j} M_{i,j}}\sum_{i,j} M_{i,j}A_{i,j} + \Delta_1\right)$$



Overview — Attention Map Consistency (AMC)

- Pretraining – from ALBEF
- Assume each sample has: V, T, M

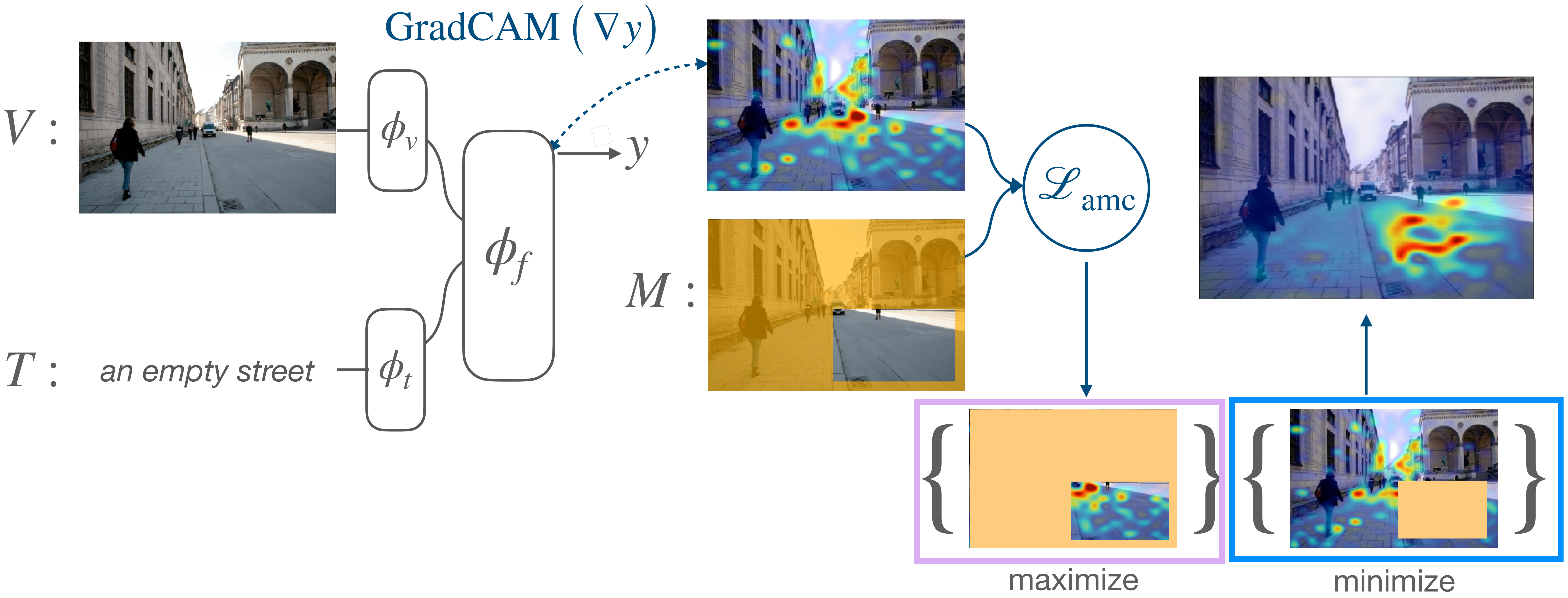
$$\mathcal{L}_{\text{mean}} = \max\left(0, \boxed{\frac{1}{\sum_{i,j}(1-M_{i,j})}\sum_{i,j}((1-M_{i,j})\,A_{i,j})} - \frac{1}{\sum_{i,j}M_{i,j}}\sum_{i,j}M_{i,j}A_{i,j} + \Delta_1\right)$$



Overview — Attention Map Consistency (AMC)

- Pretraining – from ALBEF
- Assume each sample has: V, T, M

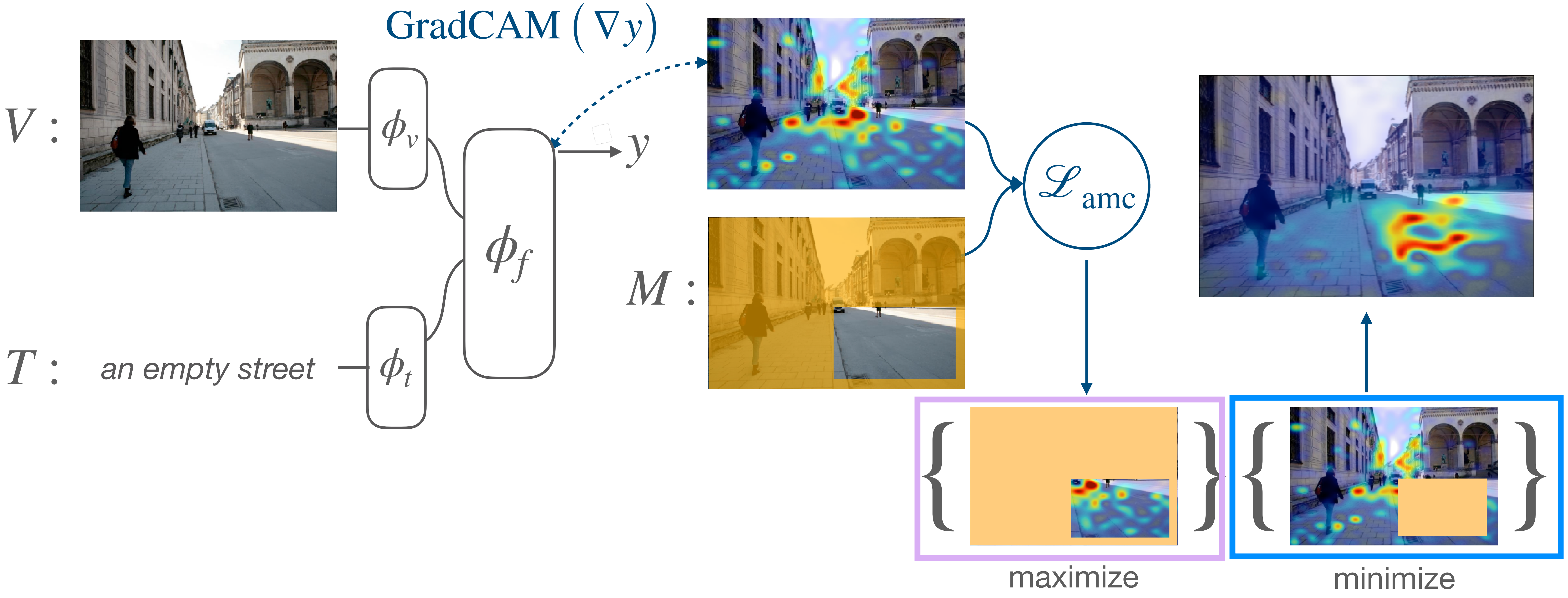
$$\mathcal{L}_{\text{mean}} = \max\left(0, \boxed{\frac{1}{\sum_{i,j}(1-M_{i,j})}\sum_{i,j}\left((1-M_{i,j})\,A_{i,j}\right)} - \boxed{\frac{1}{\sum_{i,j}M_{i,j}}\sum_{i,j}M_{i,j}A_{i,j}} + \Delta_1\right)$$



Overview — Attention Map Consistency (AMC)

- Pretraining – from ALBEF
- Assume each sample has: V, T, M

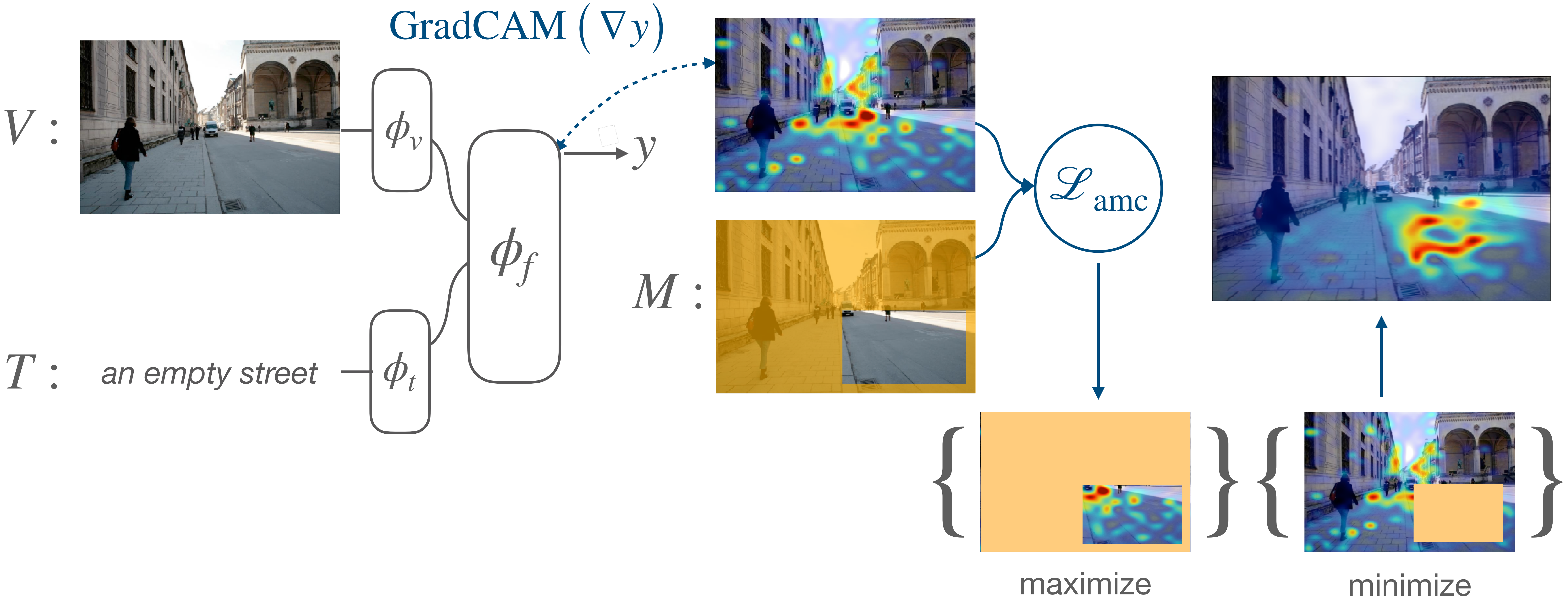
$$\mathcal{L}_{\max} = \max\left(0, \boxed{\max_{i,j}\left(\left(1 - M_{i,j}\right) A_{i,j}\right)} - \boxed{\max_{i,j} M_{i,j} A_{i,j}} + \Delta_2\right)$$



Overview — Attention Map Consistency (AMC)

- Pretraining – from ALBEF
- Assume each sample has: V, T, M

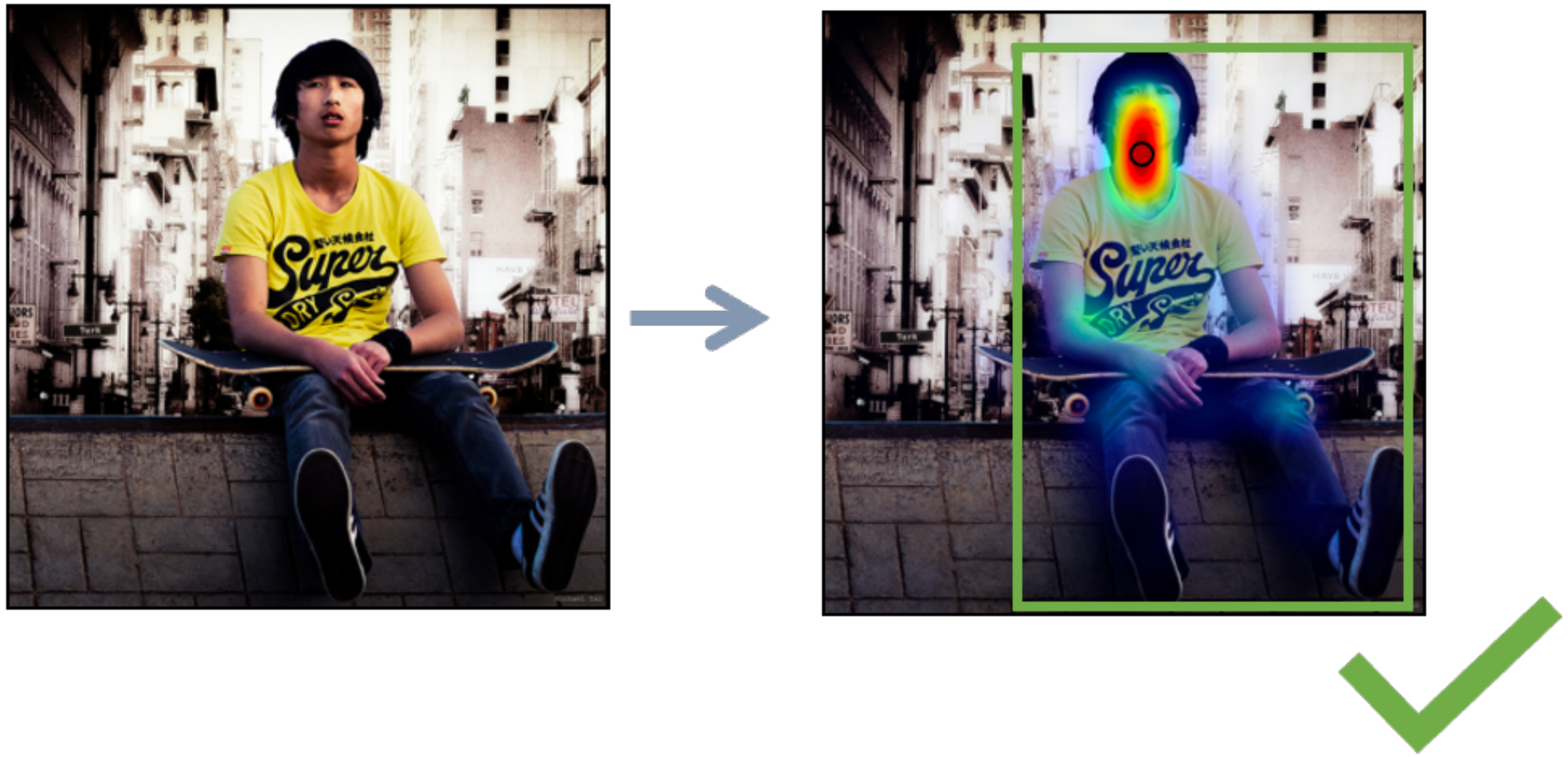
$$\mathcal{L}_{\text{amc}} = \lambda_1 \cdot \mathcal{L}_{\text{mean}} + \lambda_2 \cdot \mathcal{L}_{\text{max}}$$



Experiments

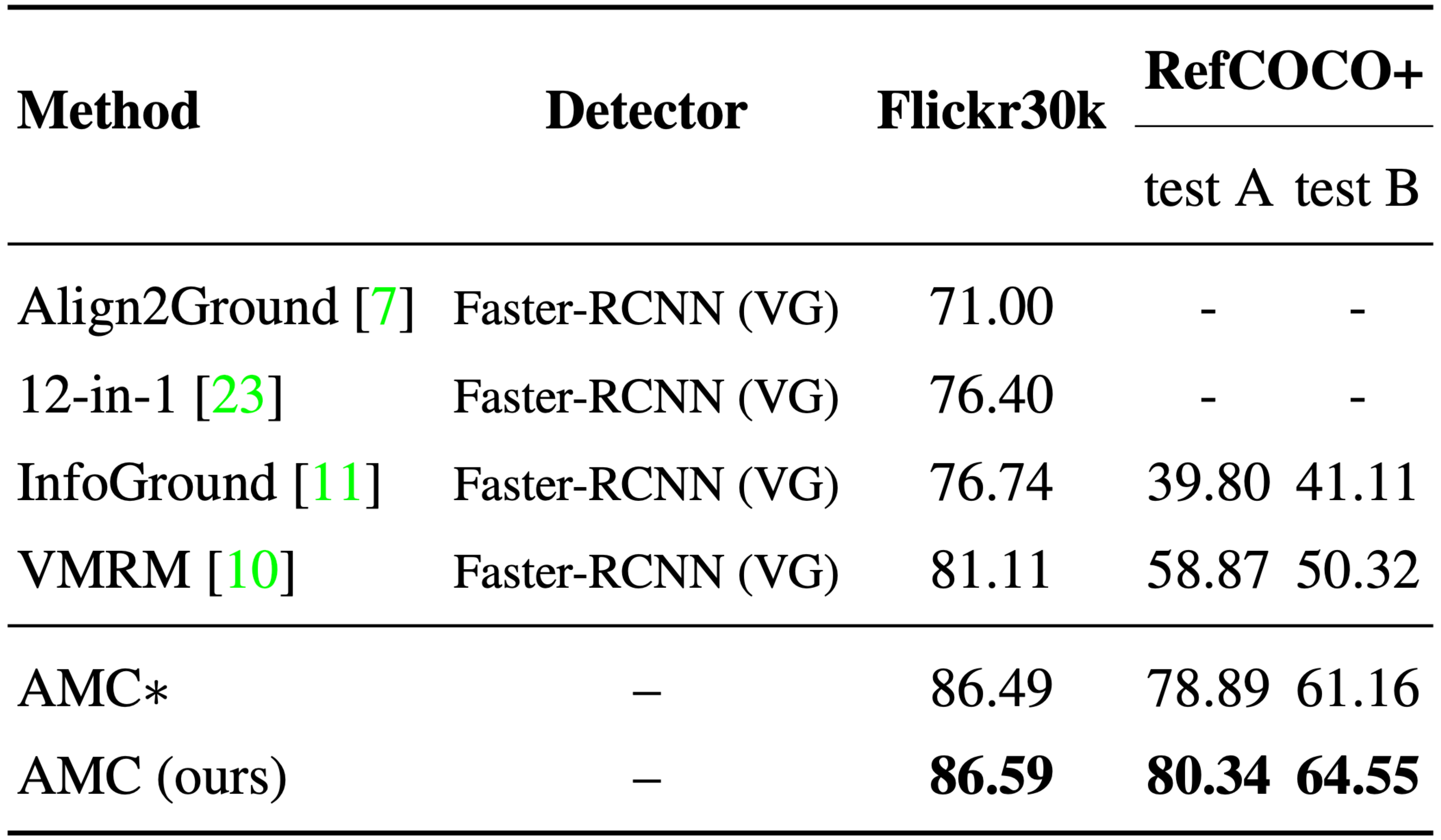
- Training Data:
  - Visual Genome
- Evaluation Data:
  - Flickr30k
  - RefCOCO+
- Evaluation metric:
  - *Pointing Game* Accuracy

A sitting asian male wearing a yellow shirt with a skateboard



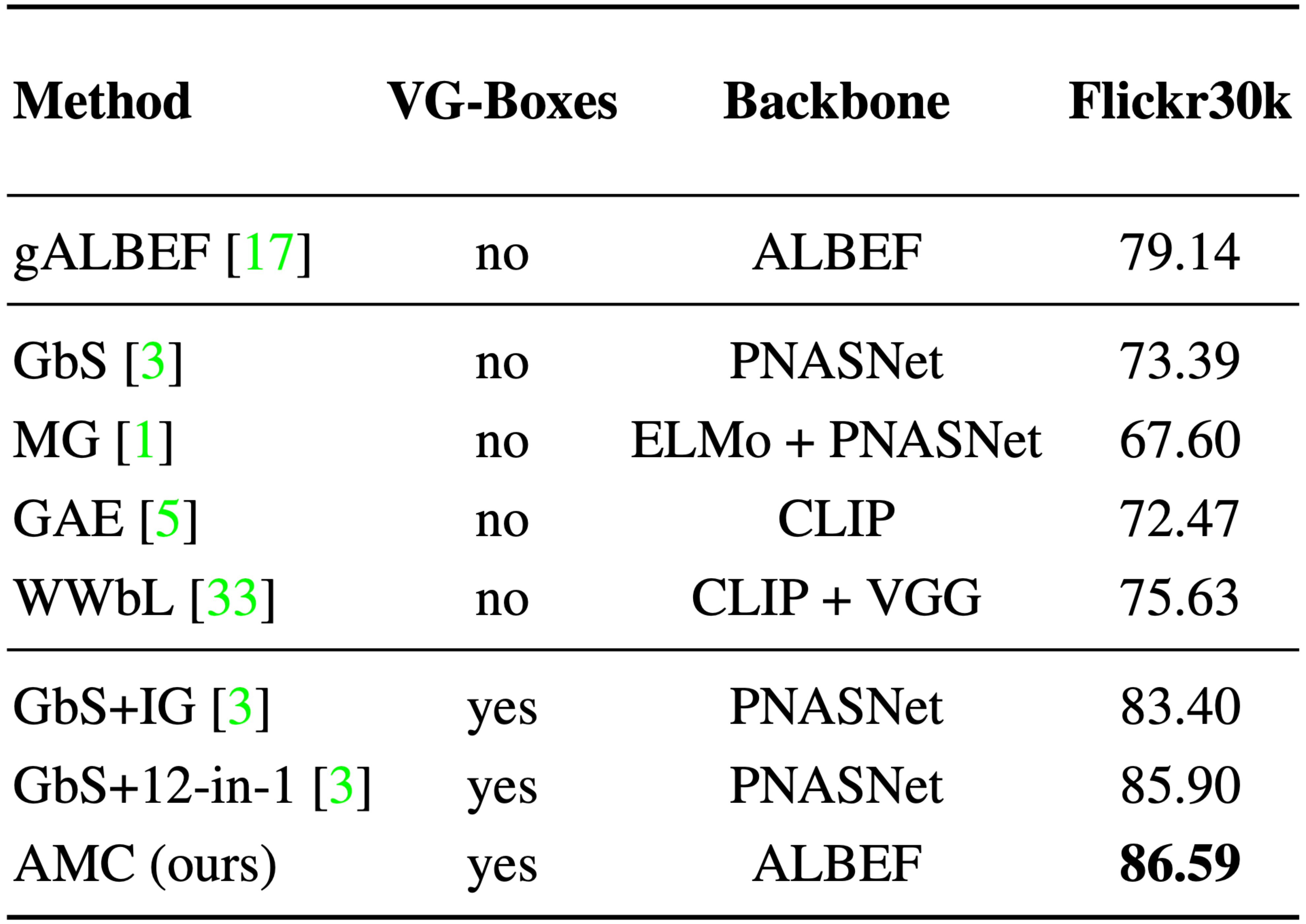
Demo: https://vislang.ai/amc

Results



Method	Detector	Flickr30k	RefCOCO+	
			test A	test B
Align2Ground [7]	Faster-RCNN (VG)	71.00	-	-
12-in-1 [23]	Faster-RCNN (VG)	76.40	-	-
InfoGround [11]	Faster-RCNN (VG)	76.74	39.80	41.11
VMRM [10]	Faster-RCNN (VG)	81.11	58.87	50.32
AMC∗	–	86.49	78.89	61.16
AMC (ours)	–	**86.59**	**80.34**	**64.55**

Table 1: Visual Grounding results using *pointing game* accuracy against methods that use different object detectors trained on Visual Genome box annotations.

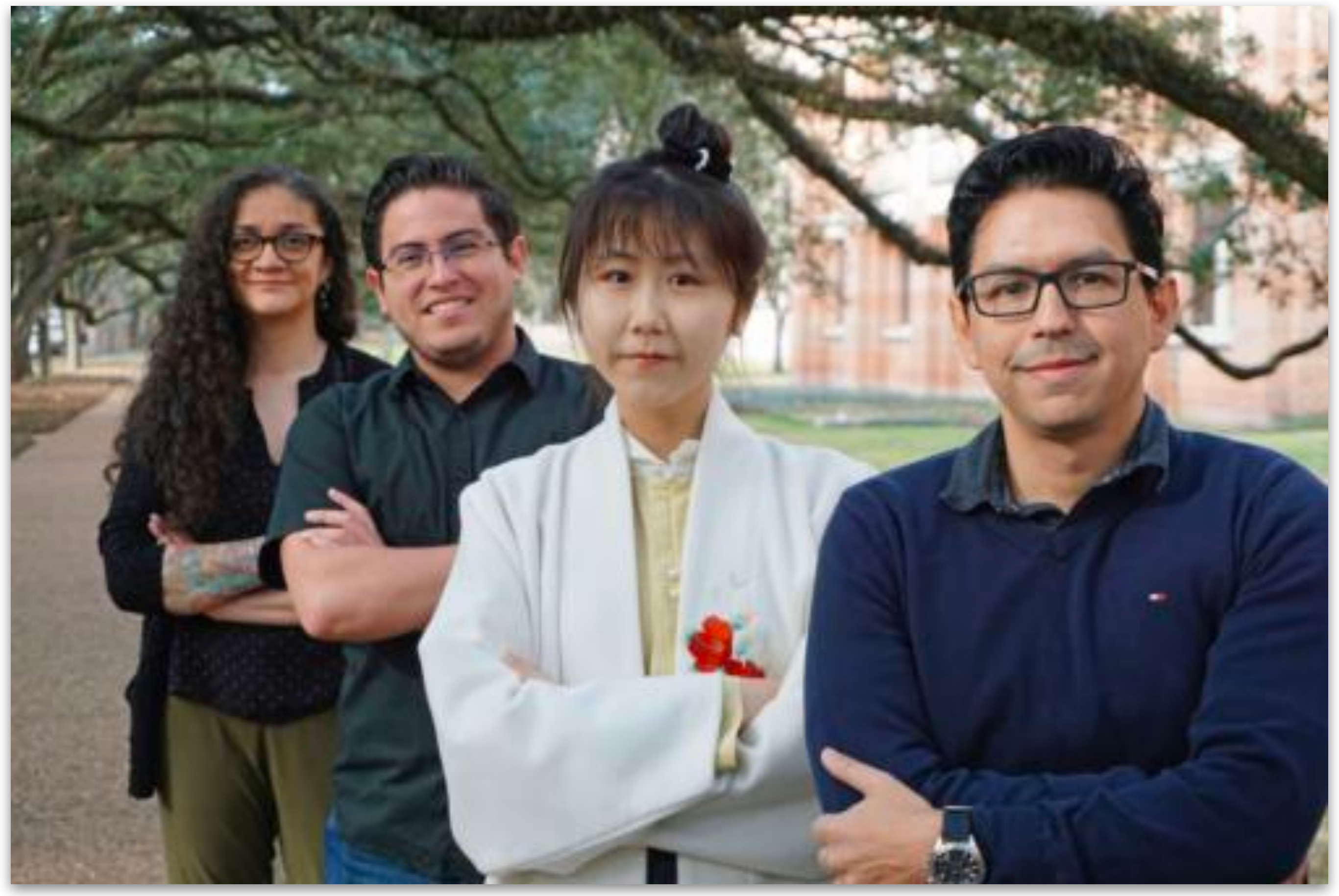


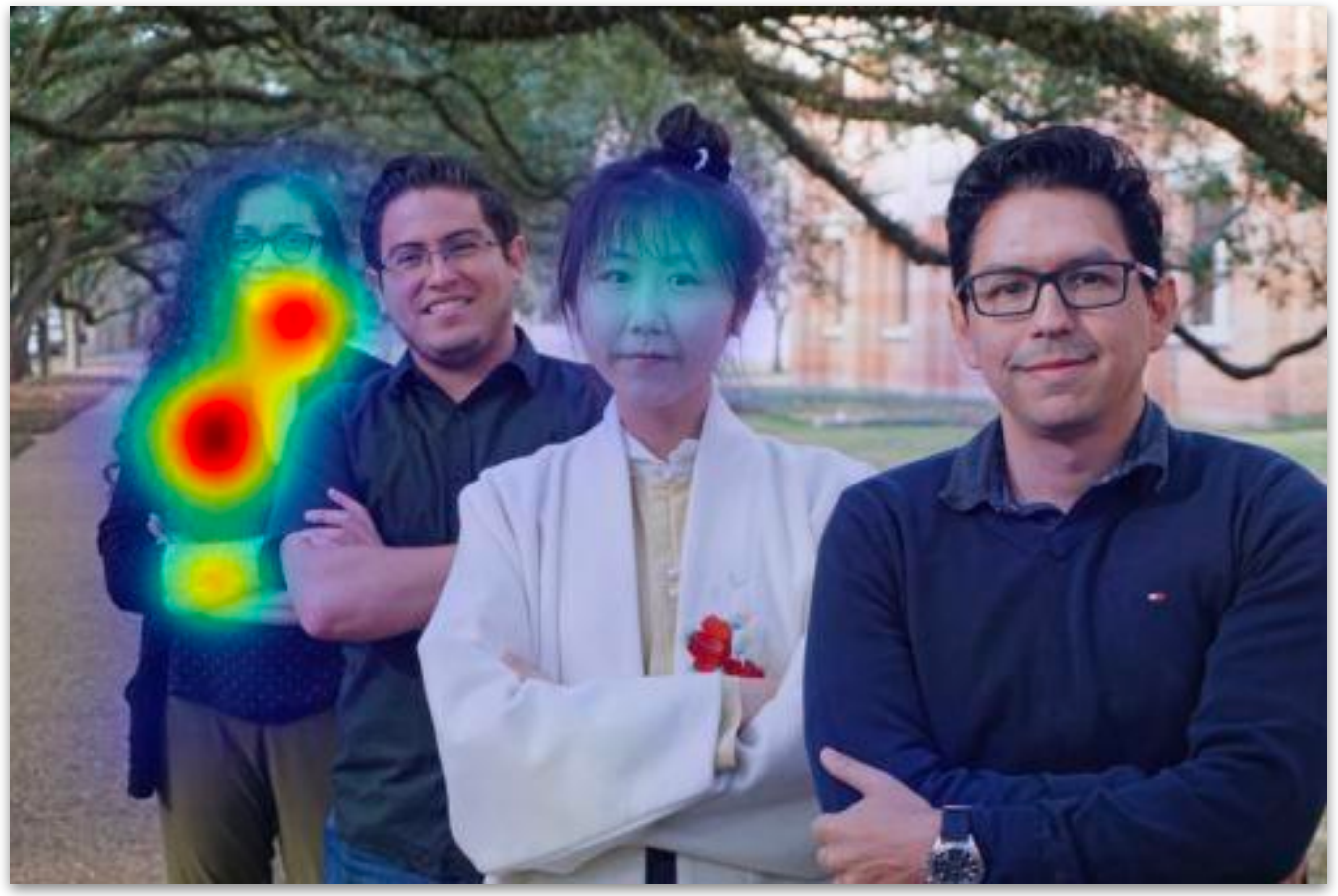
Method	VG-Boxes	Backbone	Flickr30k
gALBEF [17]	no	ALBEF	79.14
GbS [3]	no	PNASNet	73.39
MG [1]	no	ELMo + PNASNet	67.60
GAE [5]	no	CLIP	72.47
WWbL [33]	no	CLIP + VGG	75.63
GbS+IG [3]	yes	PNASNet	83.40
GbS+12-in-1 [3]	yes	PNASNet	85.90
AMC (ours)	yes	ALBEF	**86.59**

Table 2: Visual Grounding results using *pointing game* accuracy against methods that do not use object detectors or Visual Genome box supervision

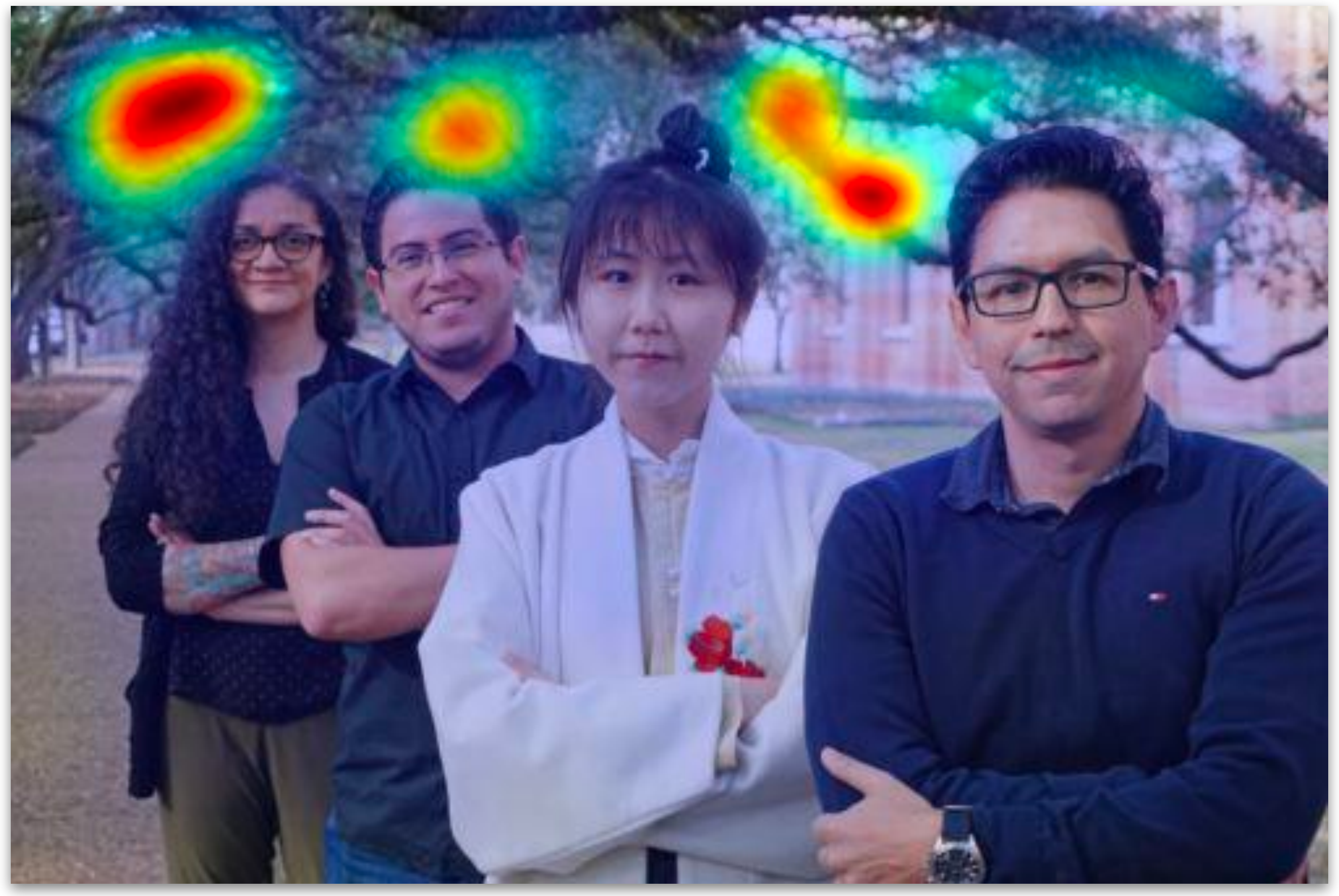
Demo: https://vislang.ai/amc

Results

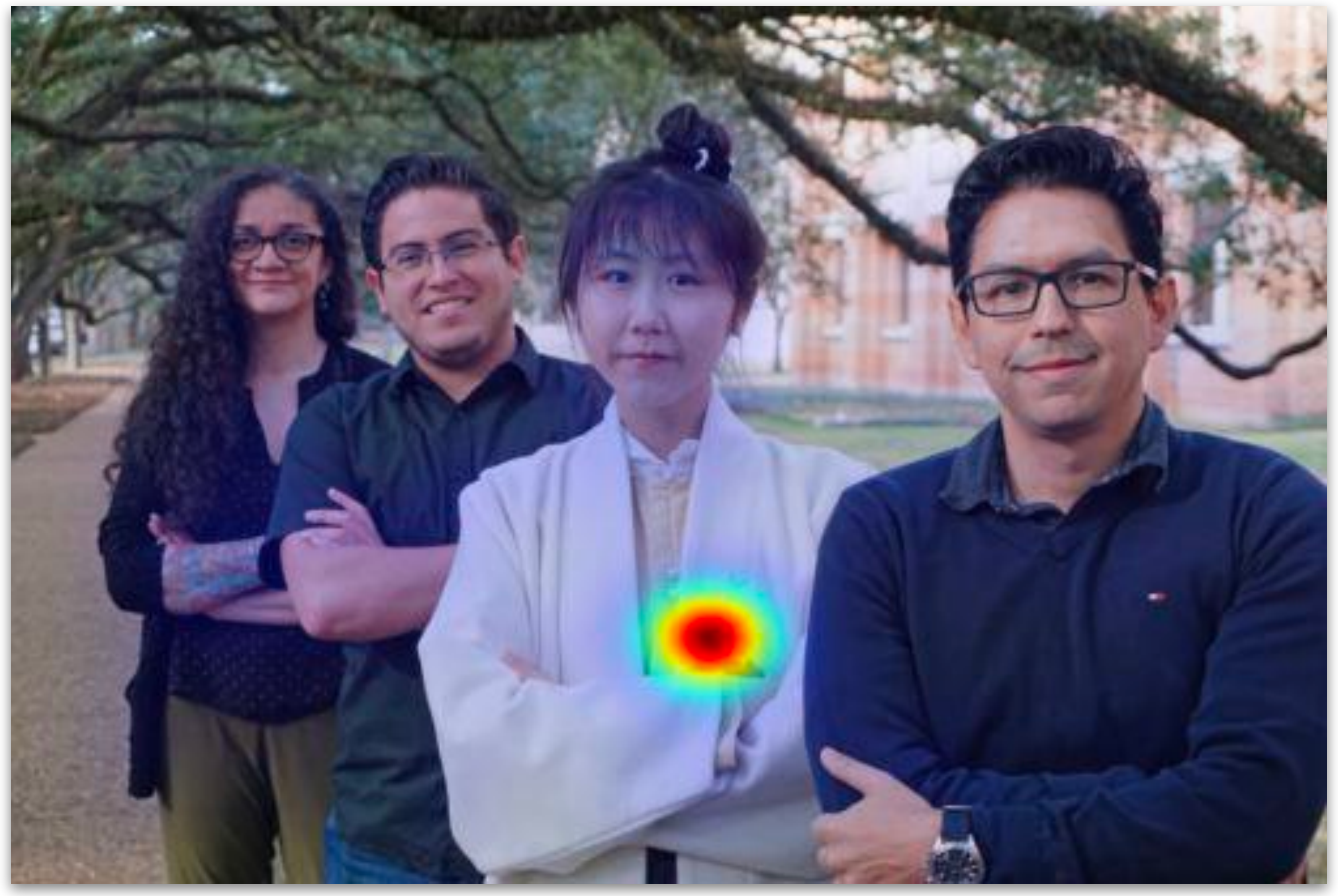




a woman with tattoo



Tree branches in the background



A red flower