PIN:
Positional Insert Unlocks Object Localisation Abilities in VLMs

MICHAEL DORKENWALD, NIMROD BARAZANI, CEES G. M. SNOEK*, YUKI M. ASANO*

CVPR'24

VLMs are great with many things, but not localisation

Prompt 1: Provide a bounding box around the cat

Prompt 2: Localise the cat in the image

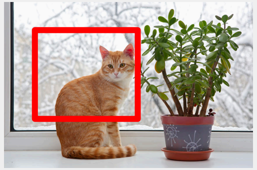
Dorkenwald, Barazani, Snoek, Asano. PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs, CVPR'24.

VLMs are great with many things, but not localisation

Prompt 1: Provide a bounding box around the cat

Prompt 2: Localise the cat in the image

P1:



***P2**: The cat in the image is sitting on the right side […]*

GPT-4V

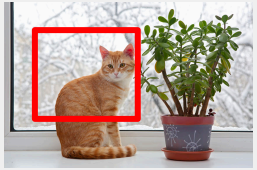
Dorkenwald, Barazani, Snoek, Asano. PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs, CVPR'24.

VLMs are great with many things, but not localisation

Prompt 1: Provide a bounding box around the cat

Prompt 2: Localise the cat in the image

P1:



***P2**: The cat in the image is sitting on the right side [...]*

GPT-4V

***P1**: To determine the size of the room.*

***P2**: The cat in the image is localised in the image.*

🦩 **OpenFlamingo**

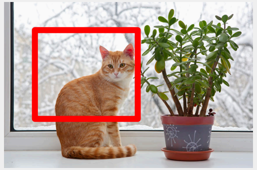
Dorkenwald, Barazani, Snoek, Asano. PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs, CVPR'24.

VLMs are great with many things, but not localisation

Prompt 1: Provide a bounding box around the cat

Prompt 2: Localise the cat in the image

P1:



P2*: The cat in the image is sitting on the right side […]*

GPT-4V

P1*: To determine the size of the room.*

P2*: The cat in the image is localised in the image.*

OpenFlamingo

P1*: Cats are not fond of being confined in a small space.*

P2*: Yes, you can do that*

FROMAGe

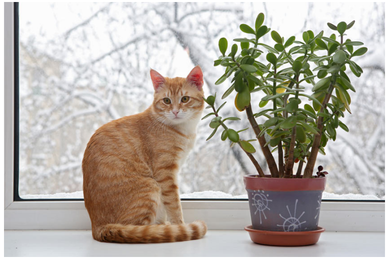
UNIVERSITY OF AMSTERDAM

Dorkenwald, Barazani, Snoek, Asano. PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs, CVPR'24.

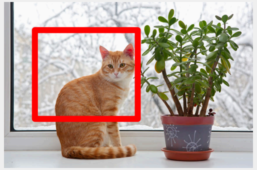
VLMs are great with many things, but not localisation

Prompt 1: Provide a bounding box around the cat

Prompt 2: Localise the cat in the image



P1:



P2*: The cat in the image is sitting on the right side […]*

GPT-4V

P1*: To determine the size of the room.*

P2*: The cat in the image is localised in the image.*

OpenFlamingo

P1*: Cats are not fond of being confined in a small space.*

P2*: Yes, you can do that*

FROMAGe

P1*: Provide a bounding box around the cat's plant*

P2*: <empty string>*

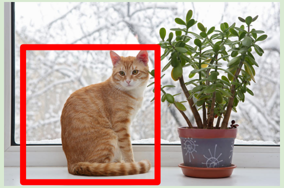
BLIP-2

Dorkenwald, Barazani, Snoek, Asano. PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs, CVPR'24.

VLMs are great with many things, but not localisation

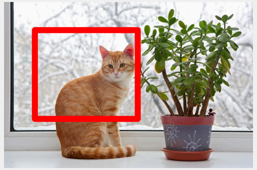
Prompt 1: Provide a bounding box around the cat

Prompt 2: Localise the cat in the image



Ours (🦩 + 🔓)

***P1*:**



***P2**: The cat in the image is sitting on the right side [...]*

GPT-4V

***P1**: To determine the size of the room.*

***P2**: The cat in the image is localised in the image.*

🦩 **OpenFlamingo**

***P1**: Cats are not fond of being confined in a small space.*

***P2**: Yes, you can do that*

FROMAGe

***P1**: Provide a bounding box around the cat's plant*

***P2**: <empty string>*

BLIP-2

Dorkenwald, Barazani, Snoek, Asano. PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs, CVPR'24.

Our approach

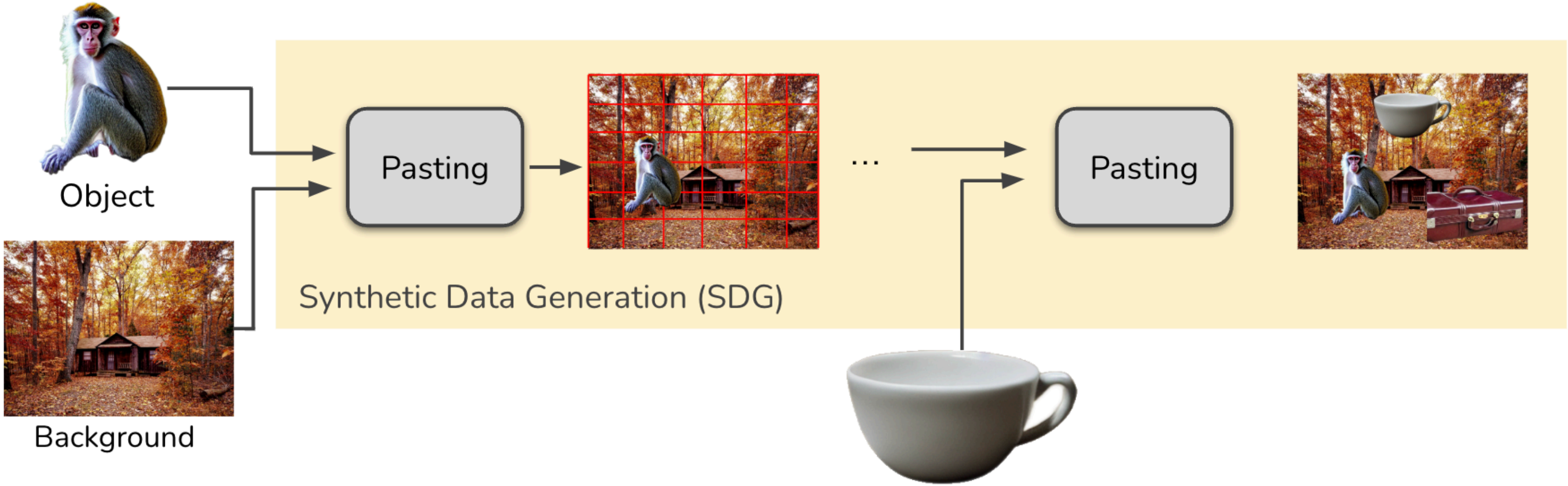
frozen VLM, e.g. Flamingo

Positional Insert (PIN) module

Synthetic, unlabelled data

Dorkenwald, Barazani, Snoek, Asano. PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs, CVPR'24.

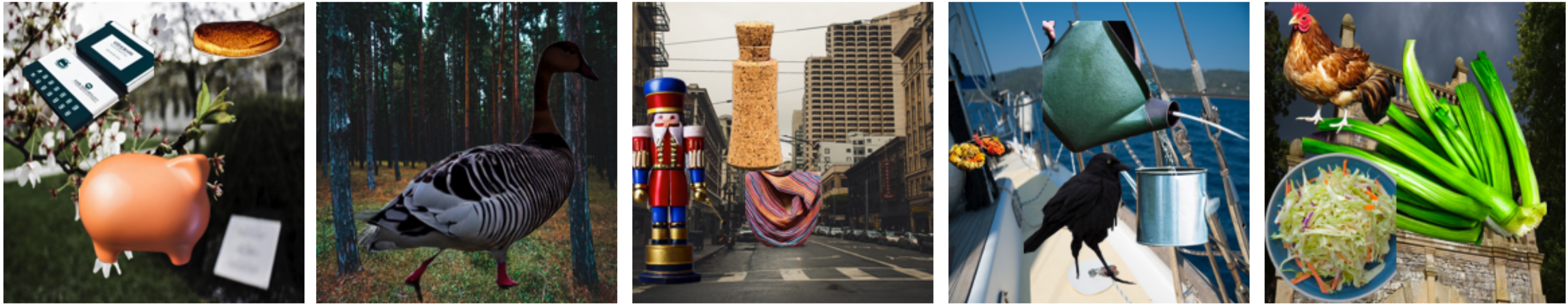
Synthetic data generation



- **Self-Supervision Signal**: Location is known via pasting
- **Avoid collapse**: Pasting multiple objects

Zhao et al. X-Paste: Revisiting Scalable Copy-Paste for Instance Segmentation using CLIP and StableDiffusion. ICML 2023

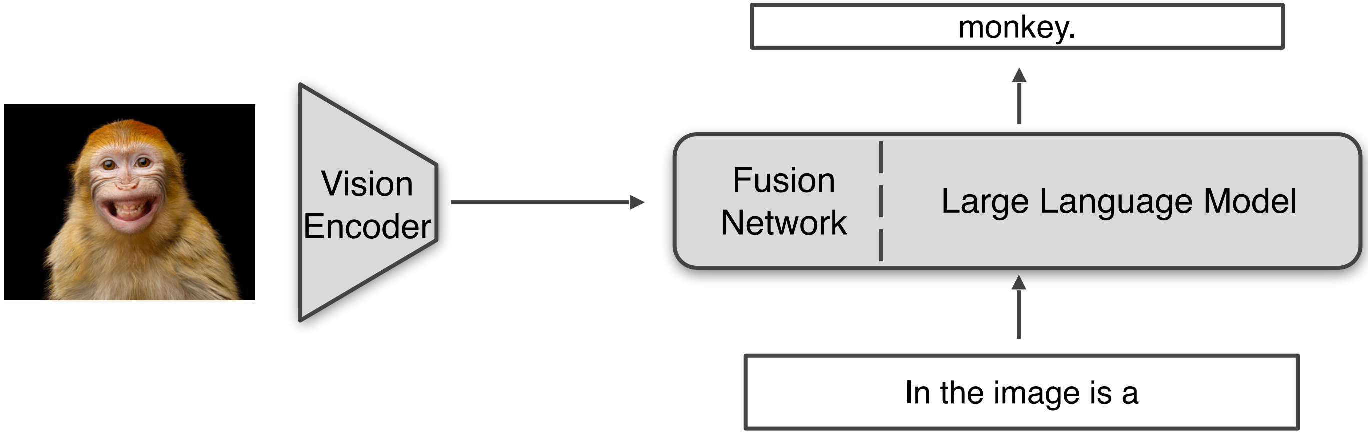
Example generated data



- Non-realism is **not** an issue, as vision encoder is kept completely frozen
- Pasting objects from categories that do not overlap with test data
  - Zero-shot evaluation

Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR'24.

Overview of VLMs



monkey.

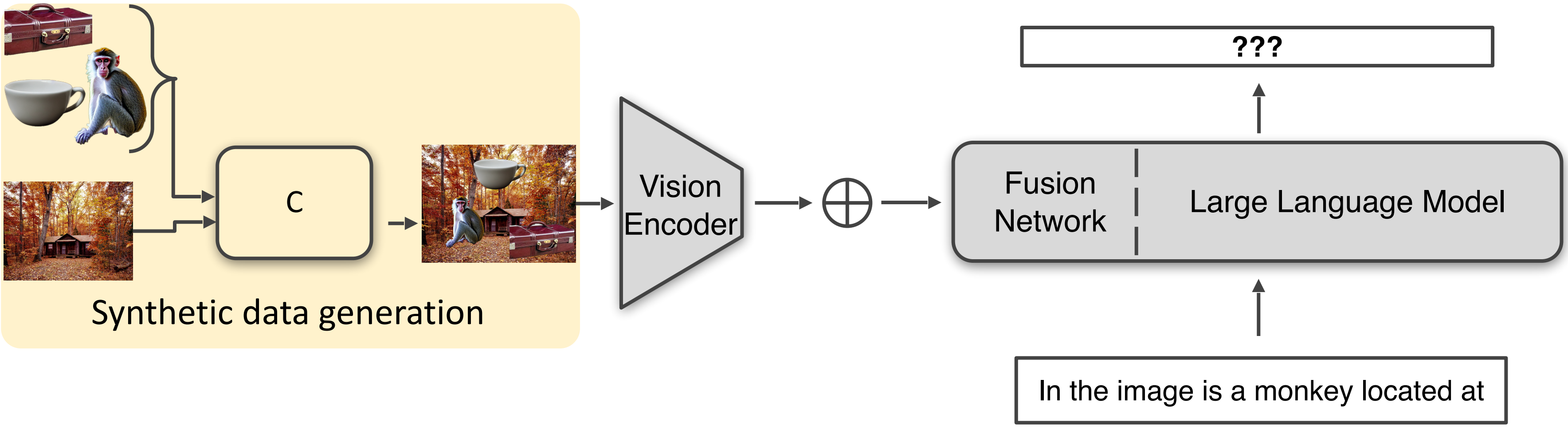
Vision Encoder

Fusion Network

Large Language Model

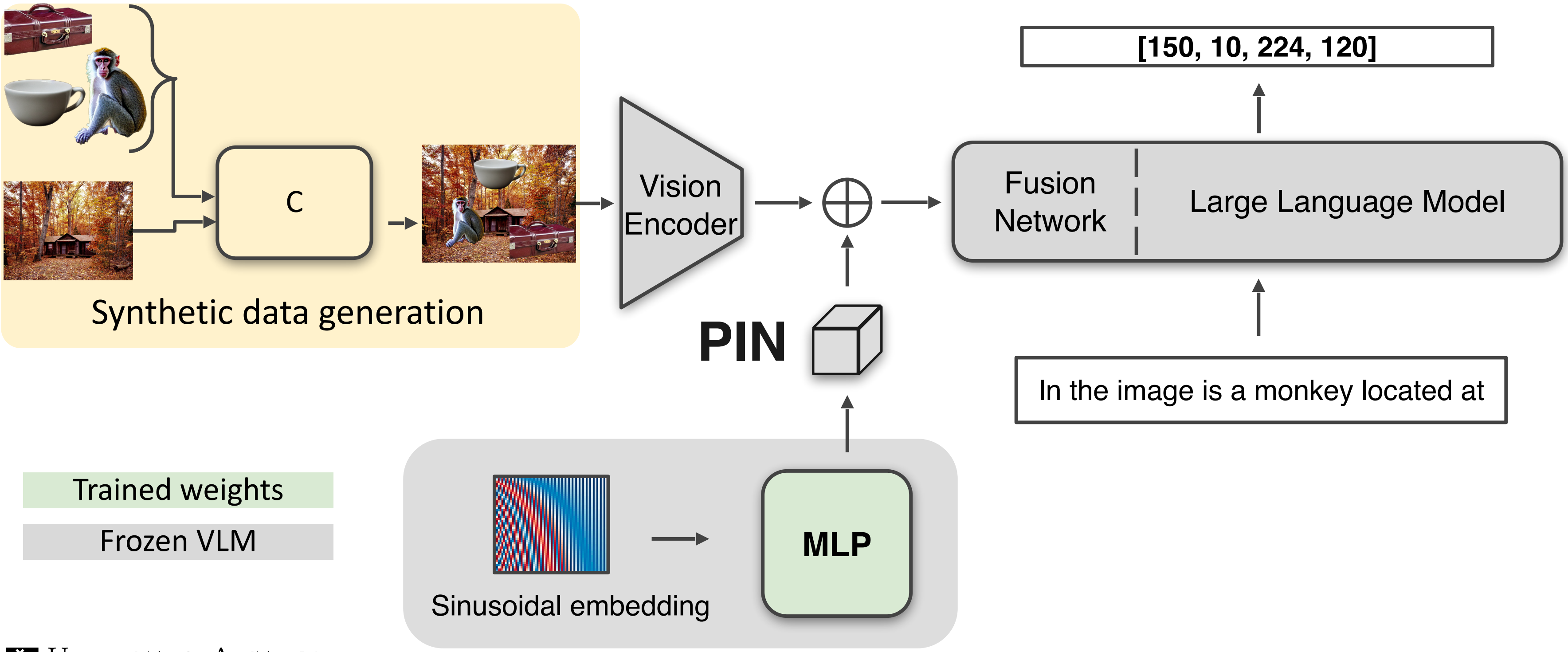
In the image is a

Feed the frozen VLM synthetic data



C
Synthetic data generation
Vision Encoder
⊕
Fusion Network
Large Language Model
???
In the image is a monkey located at

Provide spatial learning capacity via PIN



What is PIN? It's a PEFT method for VLMs

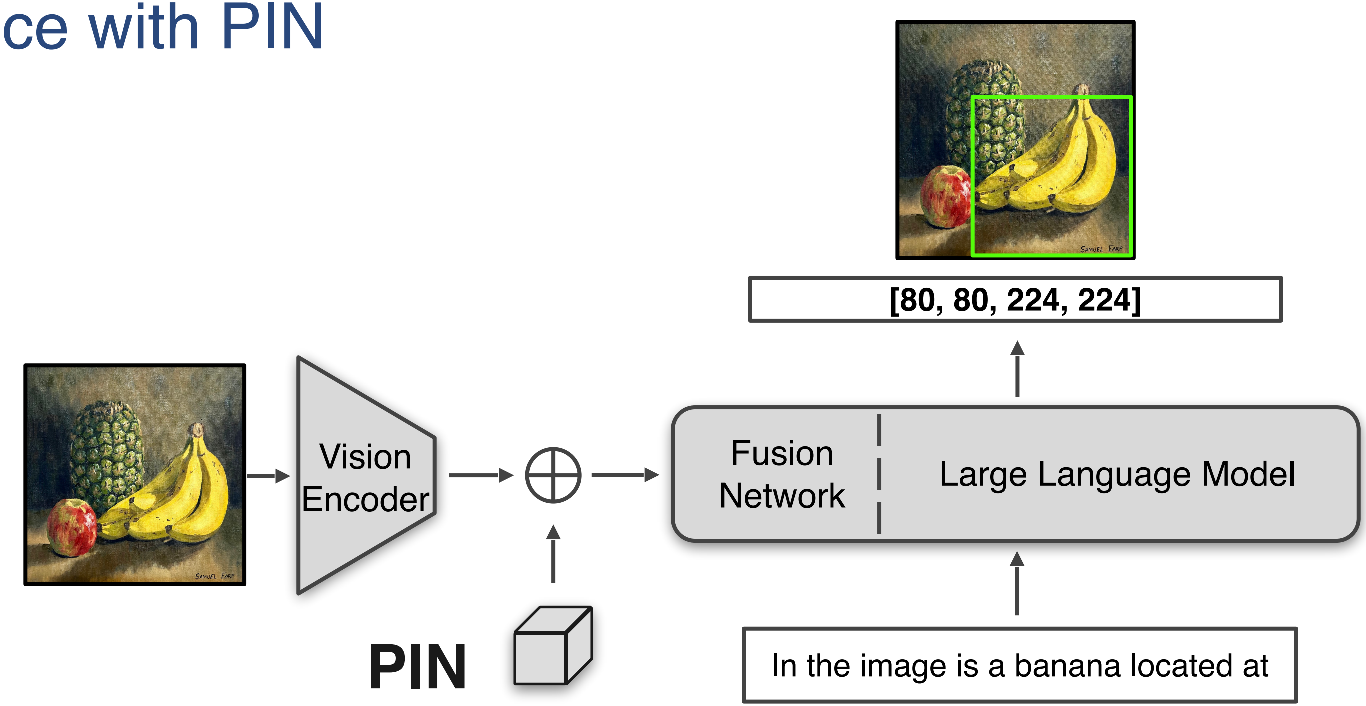
```
pos_encoding = get_sinusoid_encoding_table(n_patches=196, d_hid=64)

MLP = nn.Sequential(
    nn.Linear(64, 512),
    nn.SiLU(),
    nn.LayerNorm(512),
    nn.Linear(512, 768),
    nn.SiLU(),
    nn.LayerNorm(768),
    nn.Linear(768, 1024),
)

PIN = MLP(pos_encoding)
```

Just 10 LoC!

Inference with PIN



[80, 80, 224, 224]

Vision Encoder

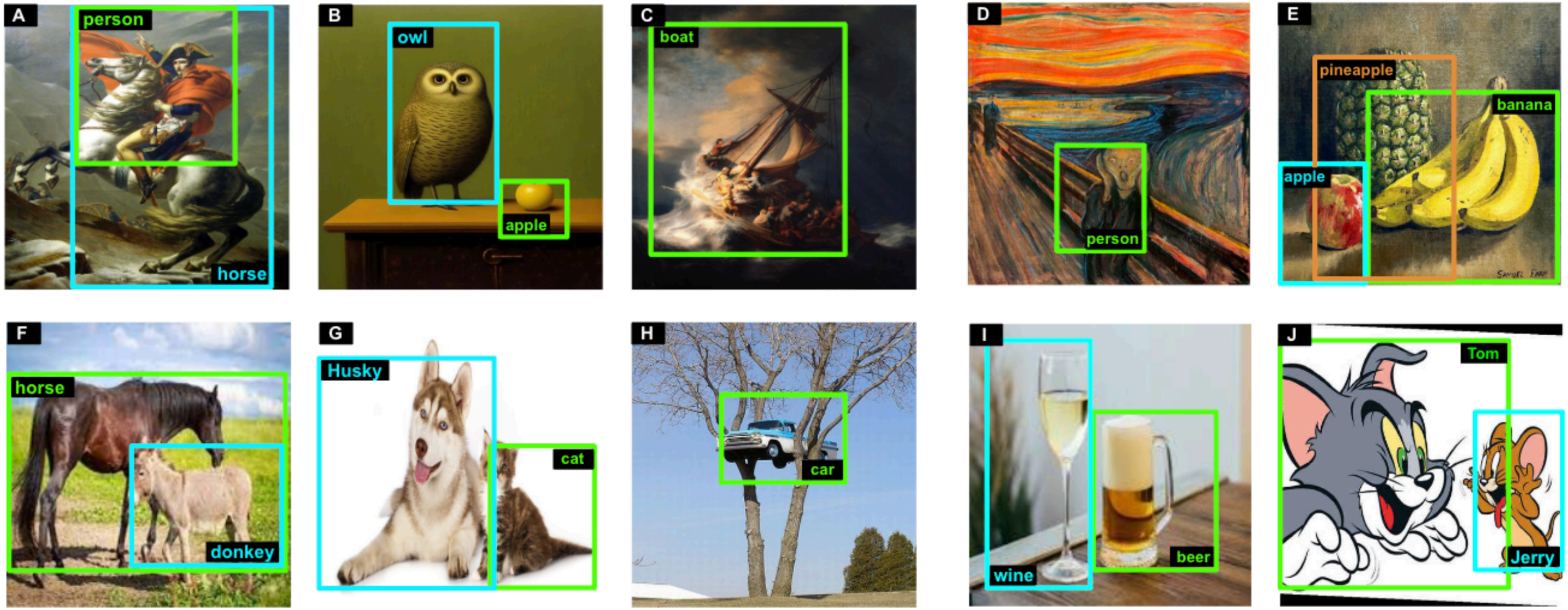
PIN

Fusion Network

Large Language Model

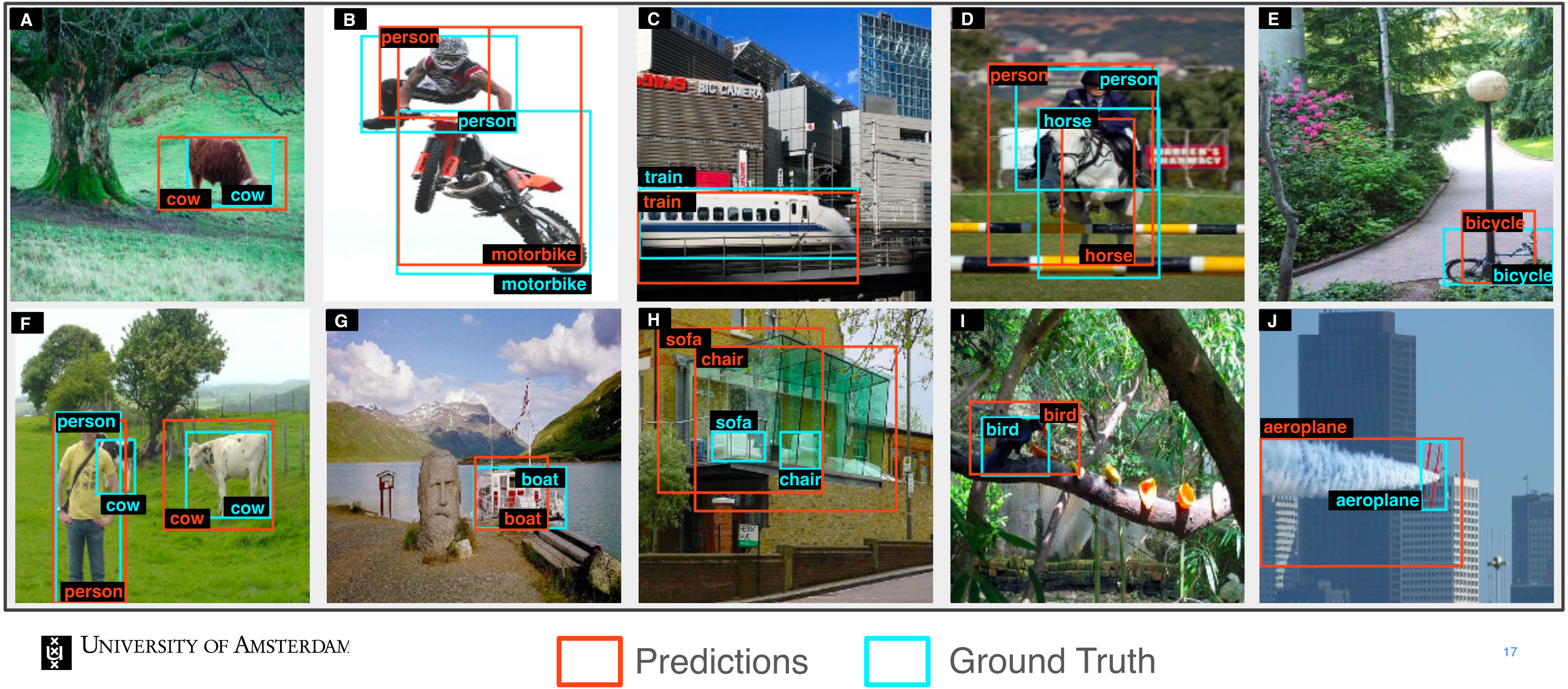
In the image is a banana located at

Results (zero-shot)

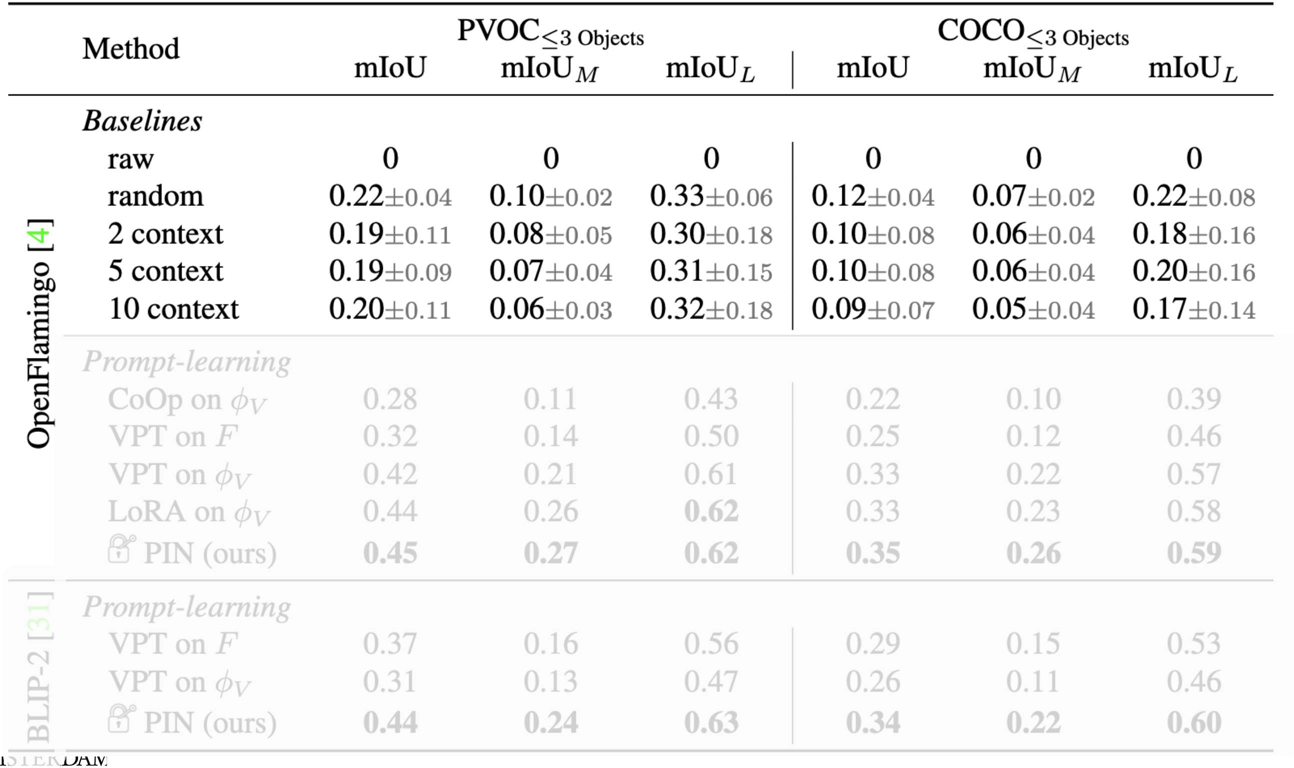


Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR'24.

Results PVOC (zero-shot)

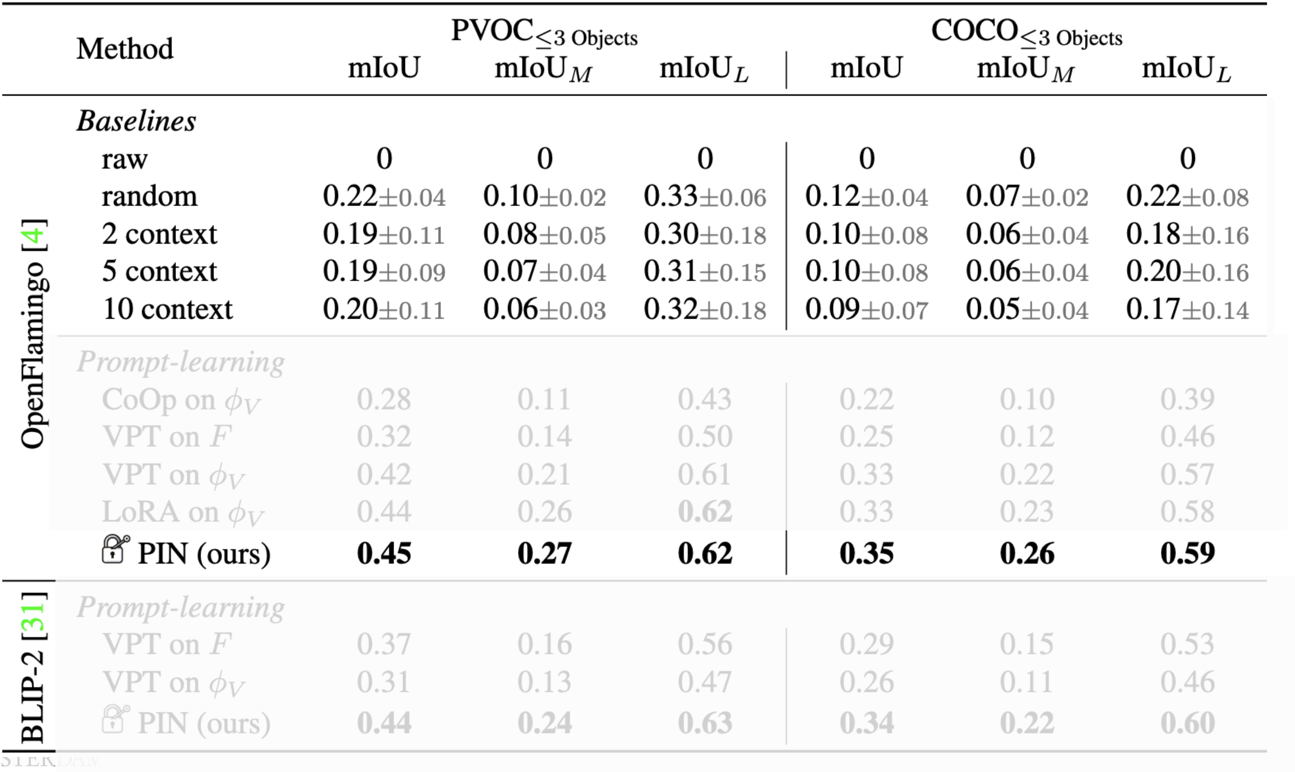


Few-shot learning doesn’t work



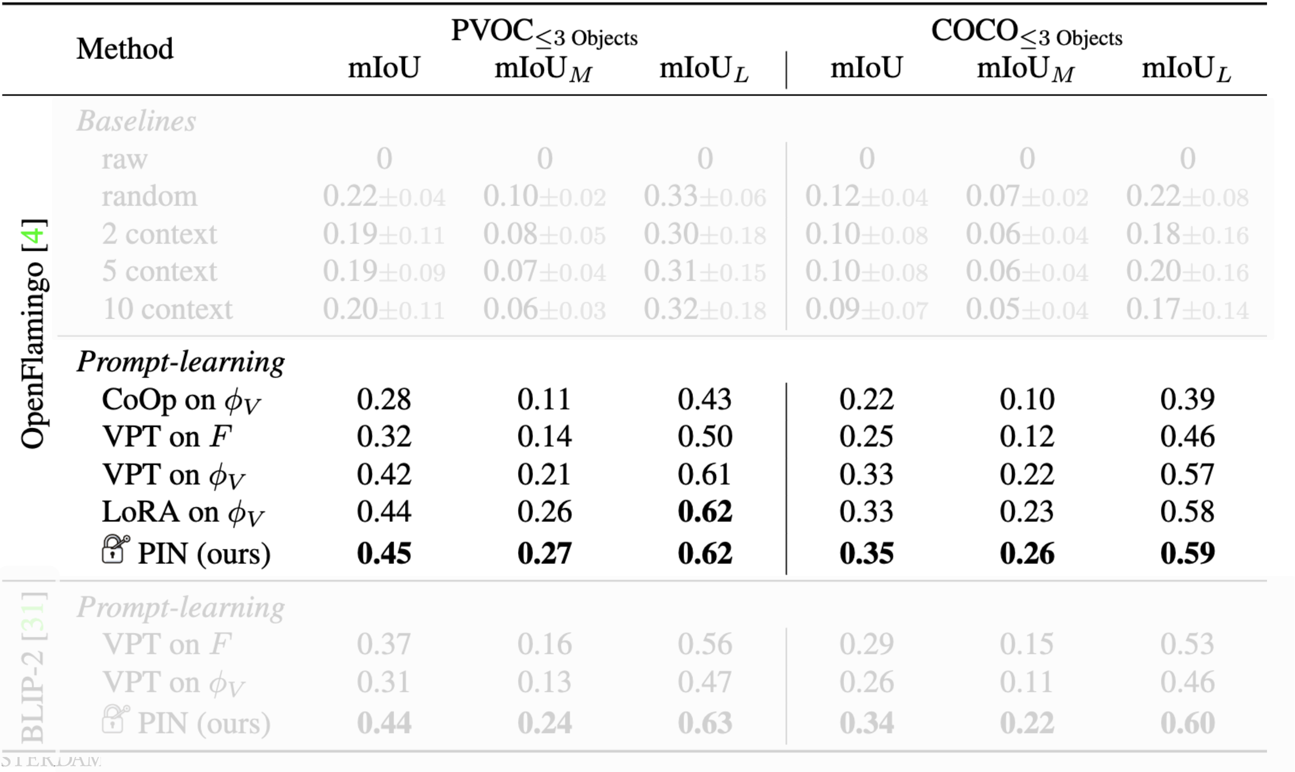
	Method	PVOC$_{\leq 3 \text{ Objects}}$ mIoU	PVOC$_{\leq 3 \text{ Objects}}$ mIoU$_M$	PVOC$_{\leq 3 \text{ Objects}}$ mIoU$_L$	COCO$_{\leq 3 \text{ Objects}}$ mIoU	COCO$_{\leq 3 \text{ Objects}}$ mIoU$_M$	COCO$_{\leq 3 \text{ Objects}}$ mIoU$_L$
OpenFlamingo [4]	*Baselines*						
	raw	0	0	0	0	0	0
	random	0.22±0.04	0.10±0.02	0.33±0.06	0.12±0.04	0.07±0.02	0.22±0.08
	2 context	0.19±0.11	0.08±0.05	0.30±0.18	0.10±0.08	0.06±0.04	0.18±0.16
	5 context	0.19±0.09	0.07±0.04	0.31±0.15	0.10±0.08	0.06±0.04	0.20±0.16
	10 context	0.20±0.11	0.06±0.03	0.32±0.18	0.09±0.07	0.05±0.04	0.17±0.14
	Prompt-learning						
	CoOp on ϕ_V	0.28	0.11	0.43	0.22	0.10	0.39
	VPT on F	0.32	0.14	0.50	0.25	0.12	0.46
	VPT on ϕ_V	0.42	0.21	0.61	0.33	0.22	0.57
	LoRA on ϕ_V	0.44	0.26	**0.62**	0.33	0.23	0.58
	🔒 PIN (ours)	**0.45**	**0.27**	**0.62**	**0.35**	**0.26**	**0.59**
BLIP-2 [31]	*Prompt-learning*						
	VPT on F	0.37	0.16	0.56	0.29	0.15	0.53
	VPT on ϕ_V	0.31	0.13	0.47	0.26	0.11	0.46
	🔒 PIN (ours)	**0.44**	**0.24**	**0.63**	**0.34**	**0.22**	**0.60**

PIN enables localisation



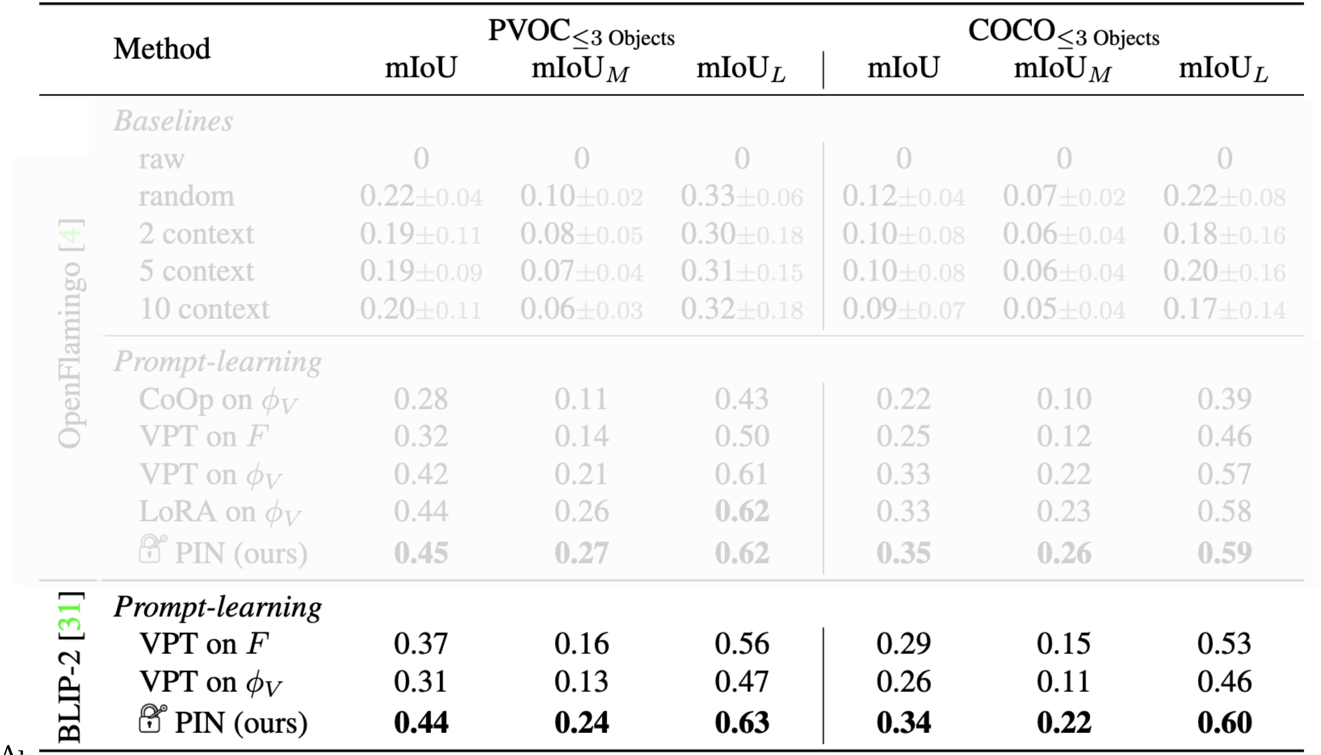
	Method	PVOC$_{\leq 3 \text{ Objects}}$ mIoU	PVOC$_{\leq 3 \text{ Objects}}$ mIoU$_M$	PVOC$_{\leq 3 \text{ Objects}}$ mIoU$_L$	COCO$_{\leq 3 \text{ Objects}}$ mIoU	COCO$_{\leq 3 \text{ Objects}}$ mIoU$_M$	COCO$_{\leq 3 \text{ Objects}}$ mIoU$_L$
OpenFlamingo [4]	*Baselines*						
	raw	0	0	0	0	0	0
	random	0.22±0.04	0.10±0.02	0.33±0.06	0.12±0.04	0.07±0.02	0.22±0.08
	2 context	0.19±0.11	0.08±0.05	0.30±0.18	0.10±0.08	0.06±0.04	0.18±0.16
	5 context	0.19±0.09	0.07±0.04	0.31±0.15	0.10±0.08	0.06±0.04	0.20±0.16
	10 context	0.20±0.11	0.06±0.03	0.32±0.18	0.09±0.07	0.05±0.04	0.17±0.14
	Prompt-learning						
	CoOp on ϕ_V	0.28	0.11	0.43	0.22	0.10	0.39
	VPT on F	0.32	0.14	0.50	0.25	0.12	0.46
	VPT on ϕ_V	0.42	0.21	0.61	0.33	0.22	0.57
	LoRA on ϕ_V	0.44	0.26	**0.62**	0.33	0.23	0.58
	PIN (ours)	**0.45**	**0.27**	**0.62**	**0.35**	**0.26**	**0.59**
BLIP-2 [31]	*Prompt-learning*						
	VPT on F	0.37	0.16	0.56	0.29	0.15	0.53
	VPT on ϕ_V	0.31	0.13	0.47	0.26	0.11	0.46
	PIN (ours)	**0.44**	**0.24**	**0.63**	**0.34**	**0.22**	**0.60**

PIN outperforms other PEFT methods



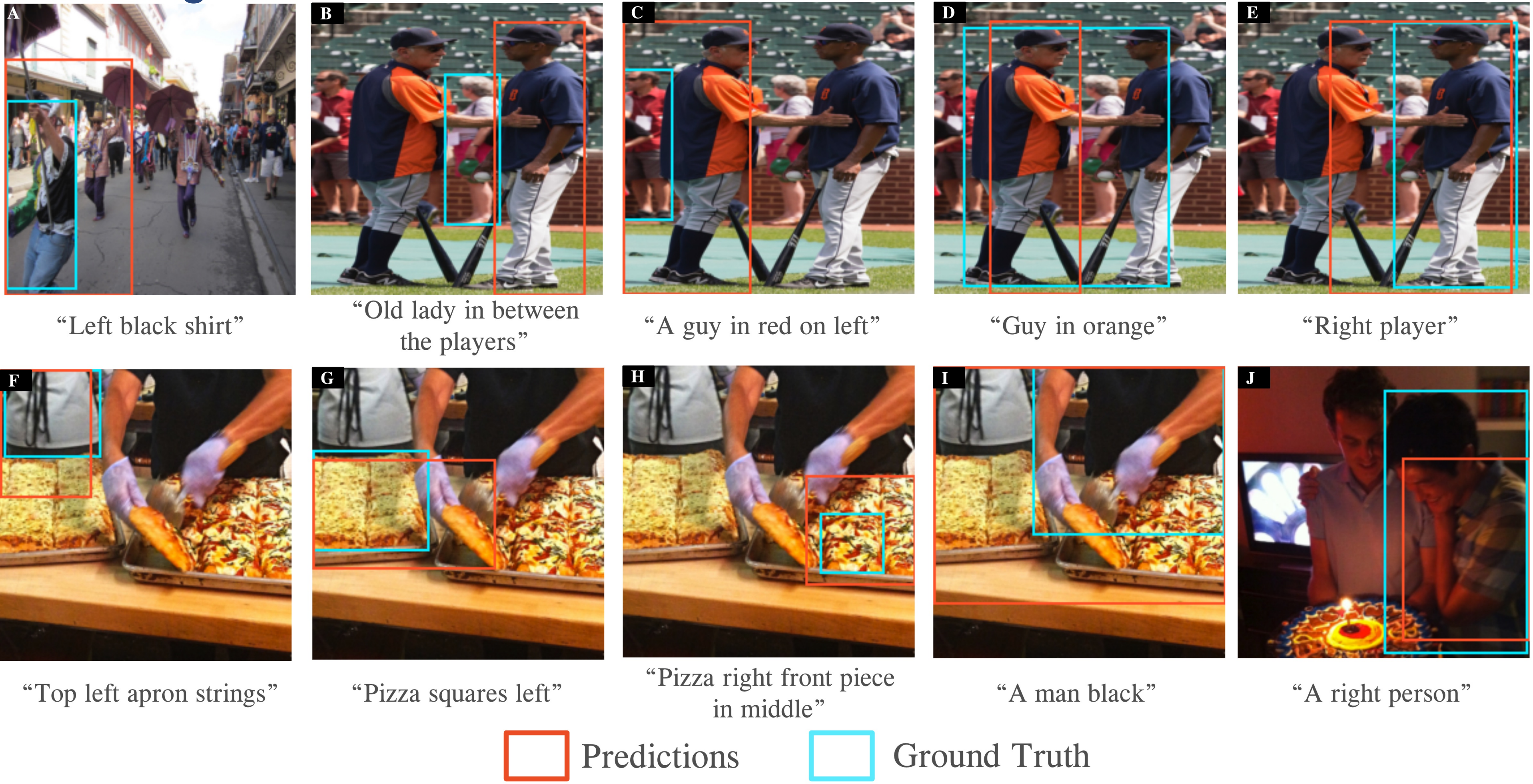
	Method	PVOC$_{\leq 3 \text{ Objects}}$			COCO$_{\leq 3 \text{ Objects}}$		
		mIoU	mIoU$_M$	mIoU$_L$	mIoU	mIoU$_M$	mIoU$_L$
OpenFlamingo [4]	*Baselines*						
	raw	0	0	0	0	0	0
	random	0.22±0.04	0.10±0.02	0.33±0.06	0.12±0.04	0.07±0.02	0.22±0.08
	2 context	0.19±0.11	0.08±0.05	0.30±0.18	0.10±0.08	0.06±0.04	0.18±0.16
	5 context	0.19±0.09	0.07±0.04	0.31±0.15	0.10±0.08	0.06±0.04	0.20±0.16
	10 context	0.20±0.11	0.06±0.03	0.32±0.18	0.09±0.07	0.05±0.04	0.17±0.14
	Prompt-learning						
	CoOp on ϕ_V	0.28	0.11	0.43	0.22	0.10	0.39
	VPT on F	0.32	0.14	0.50	0.25	0.12	0.46
	VPT on ϕ_V	0.42	0.21	0.61	0.33	0.22	0.57
	LoRA on ϕ_V	0.44	0.26	**0.62**	0.33	0.23	0.58
	🔐 PIN (ours)	**0.45**	**0.27**	**0.62**	**0.35**	**0.26**	**0.59**
BLIP-2 [31]	*Prompt-learning*						
	VPT on F	0.37	0.16	0.56	0.29	0.15	0.53
	VPT on ϕ_V	0.31	0.13	0.47	0.26	0.11	0.46
	🔐 PIN (ours)	**0.44**	**0.24**	**0.63**	**0.34**	**0.22**	**0.60**

PIN works on other VLMs too



	Method	PVOC$_{\leq 3 \text{ Objects}}$			COCO$_{\leq 3 \text{ Objects}}$		
		mIoU	mIoU$_M$	mIoU$_L$	mIoU	mIoU$_M$	mIoU$_L$
OpenFlamingo [4]	*Baselines*						
	raw	0	0	0	0	0	0
	random	0.22±0.04	0.10±0.02	0.33±0.06	0.12±0.04	0.07±0.02	0.22±0.08
	2 context	0.19±0.11	0.08±0.05	0.30±0.18	0.10±0.08	0.06±0.04	0.18±0.16
	5 context	0.19±0.09	0.07±0.04	0.31±0.15	0.10±0.08	0.06±0.04	0.20±0.16
	10 context	0.20±0.11	0.06±0.03	0.32±0.18	0.09±0.07	0.05±0.04	0.17±0.14
	Prompt-learning						
	CoOp on ϕ_V	0.28	0.11	0.43	0.22	0.10	0.39
	VPT on F	0.32	0.14	0.50	0.25	0.12	0.46
	VPT on ϕ_V	0.42	0.21	0.61	0.33	0.22	0.57
	LoRA on ϕ_V	0.44	0.26	**0.62**	0.33	0.23	0.58
	PIN (ours)	**0.45**	**0.27**	**0.62**	**0.35**	**0.26**	**0.59**
BLIP-2 [31]	*Prompt-learning*						
	VPT on F	0.37	0.16	0.56	0.29	0.15	0.53
	VPT on ϕ_V	0.31	0.13	0.47	0.26	0.11	0.46
	PIN (ours)	**0.44**	**0.24**	**0.63**	**0.34**	**0.22**	**0.60**

With slight modification, can work on RefCOCO.



A: "Left black shirt"

B: "Old lady in between the players"

C: "A guy in red on left"

D: "Guy in orange"

E: "Right player"

F: "Top left apron strings"

G: "Pizza squares left"

H: "Pizza right front piece in middle"

I: "A man black"

J: "A right person"

Thank you!