LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

ICLR 2025

Amaia Cardiel

Eloi Zablocki

Elias Ramzi

Oriane Siméoni

Matthieu Cord

valeo.ai

UGA Université Grenoble Alpes

SORBONNE UNIVERSITÉ

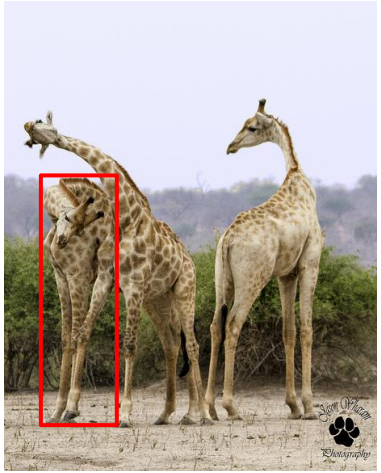
ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

The REC Task

- **Task**: Given a text query, find the **single bounding box** in an image around the **described object**
- **Main challenge: spatial and semantic understanding** to distinguish described object among **several distractor objects**



RefCOCO [1] (~ 3.5 words / query)
Query: "Giraffe on left"

RefCOCO+ [1] (~ 3.5 words / query)
Query: "Giraffe head down"

RefCOCOg [2] (~ 8.3 words / query)
Query: "An adult giraffe scratching its back with its horn"

Classic benchmarks

Talk2Car [3] (~ 11 words / query)
Query: "You can park up ahead behind the silver car, next to that lamppost with the orange sign on it."

RefLoCo [4] (~ 84.6 words / query)
Query: "The person is outfitted in a distinctive black and yellow full-body uniform, with the "DEWALT" brand emblazoned across the chest area. A black helmet, equipped with a visor, adorns his head, and he is frozen in a dynamic action stance. His involvement with a pit crew is suggested by the act of refueling a race car, which is indicated by the sizeable red fuel container he is deftly handling and utilizing."

More challenging datasets

[1] Kazemzadeh et al. Referitgame: Referring to objects in photographs of natural scenes. EMNLP 2014
[2] Mao et al. Generation and comprehension of unambiguous object descriptions. CVPR 2016.
[3] Deruyttere et al. Talk2Car: Taking Control of Your Self-Driving Car. EMNLP/IJCNLP 2019.
[4] Wei et al. A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era. NeurIPS 2024.

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

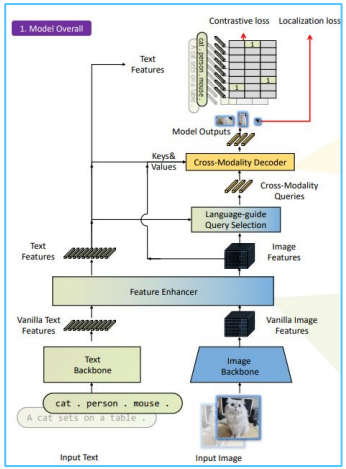
Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

ICLR 2025

VLMs & REC

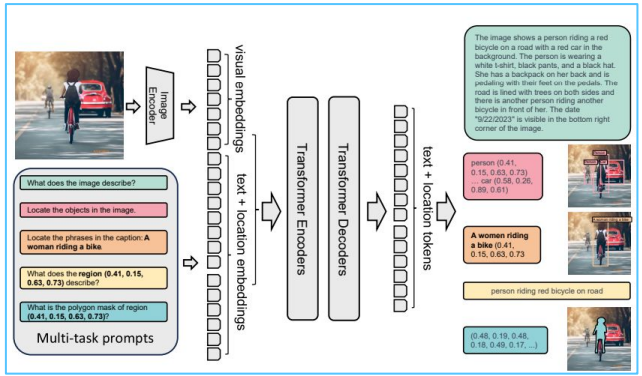
Some highly performing VLMs on REC

Grounding DINO [1]



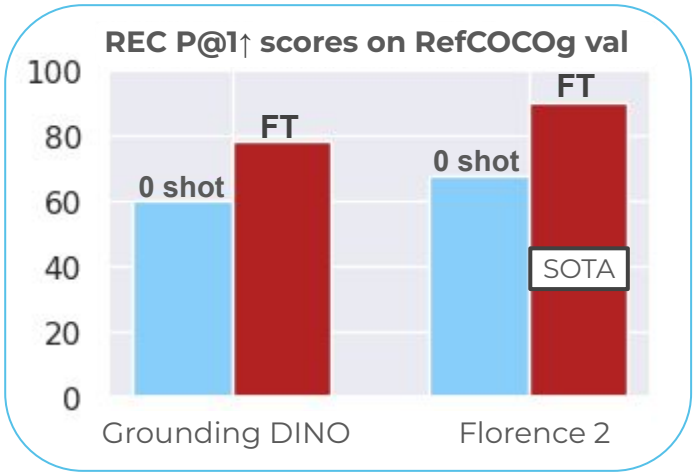
deep modality fusion

Florence 2 [2]



sequence-to-sequence modeling

VLMs excel at REC when fine-tuned



[1] Liu et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. ECCV 2024.
[2] Xiao et al. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. CVPR 2024.

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

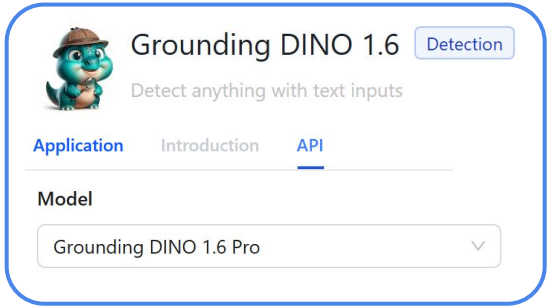
Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Motivation

- **Why "Black-Box" adaptation ?**

VLMs need to be fine-tuned to be competitive on the REC task but **full fine-tuning has limitations**:

- It is costly
- It requires task & model specific knowledge
- It requires full model access
- No possible transfer to new VLMs or datasets



ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

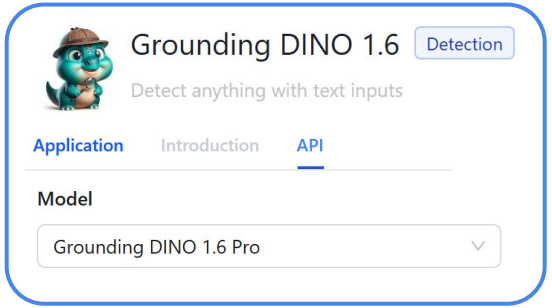
Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Motivation

- **Why "Black-Box" adaptation ?**

VLMs need to be fine-tuned to be competitive on the REC task but **full fine-tuning has limitations**:

  - It is costly
  - It requires task & model specific knowledge
  - It requires full model access
  - No possible transfer to new VLMs or datasets

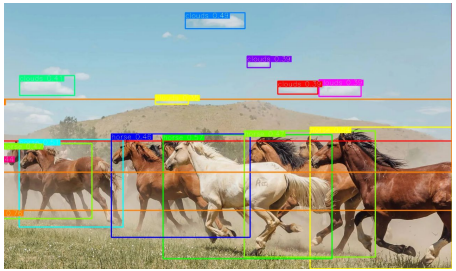


- **Why "Semantic-Aware" adaptation ?**

  - Most zero-shot **VLMs already output tight, properly labelled bounding boxes** but they fail at compositional understanding.

  - The main axis to adapt VLMs for the REC task is to **improve their semantic understanding using an LLM.**

VLMs (e.g. Grounding DINO) output tight well-labelled bounding boxes



Query: "Horse. Clouds. Grasses. Sky. Hill"

ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Method

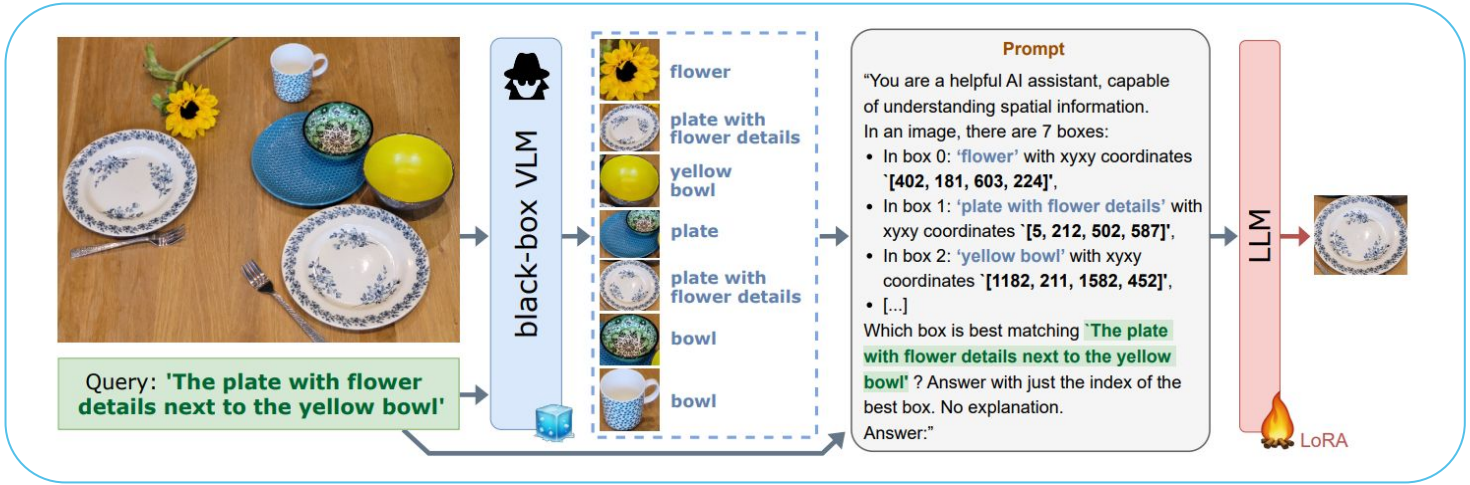


Illustration of LLM-wrapper

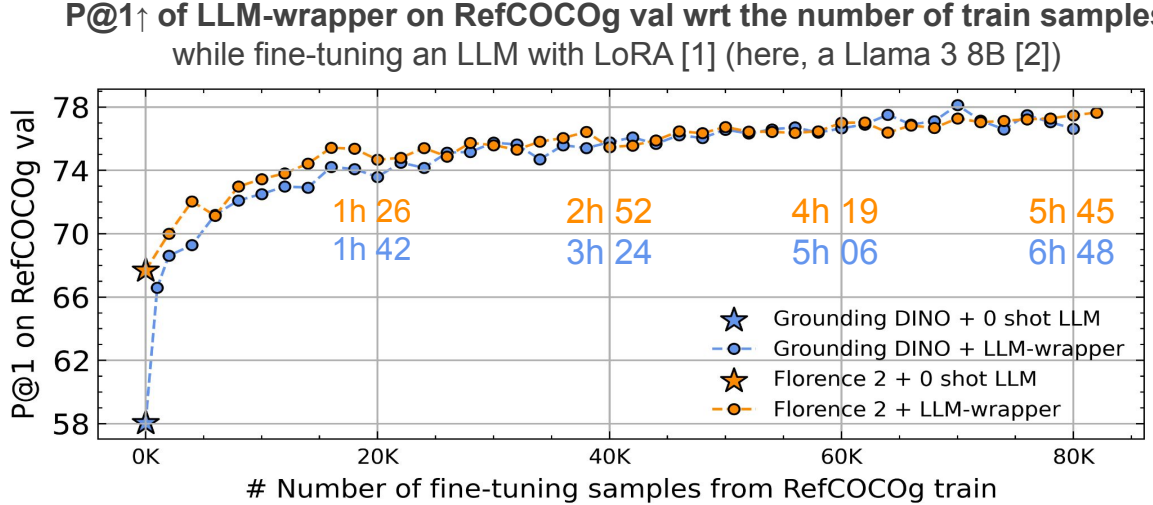
- **VLM outputs** (bounding box coordinates and labels) are **translated into natural language** to forge a prompt
- An **LLM** learns to **identify the best box among given candidates** with a LoRA fine-tuning

ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Method



Training setup

- We build training data using REC data + VLM outputs
- We permute the box proposals to avoid shortcut learning

Resulting fine-tuned LLM is robust to failure-cases

- 0% of its outputs are non-integer
- Less than 0.3% of its outputs are “out of range”

[1] Hu et al. LoRA: Low-Rank Adaptation of LLMs. ICLR 2022
[2] Grattafiori et al. The Llama 3 Herd of Models. CoRR 2407.21783, 2024.

ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Results

Main results of LLM-wrapper on REC (in P@1↑)

Adaptation	Model access	VLM	LLM	RefCOCOg		RefCOCO		RefCOCO+		Talk2Car	
				val-umd	test-umd	val-unc	test-unc	val-unc	test-unc	val	test
∅ (zero-shot)		GD (T)		60.09	59.32	50.69	50.94	51.65	51.79	55.37	58.44
Fine-tuning	*White-box*	*GD (B)*		*78.51*	*77.99*	*83.86*	*84.12*	*73.46*	*73.46*	N/A	N/A
LLM-wrapper	Black-box	GD (T)	Mixtral	77.57 ↑17.5	77.05 ↑17.7	74.61↑23.9	73.46↑22.5	60.32↑8.7	60.08↑8.3	64.75 ↑9.4	67.14 ↑8.7
LLM-wrapper	Black-box	GD (T)	Llama3	78.12 ↑18.0	77.36 ↑18.0	74.78↑24.1	73.98↑23.0	64.18↑12.5	63.82↑12.0	65.95 ↑10.6	68.61 ↑10.2
∅ (zero-shot)		Flo2		67.91	66.16	55.94	57.21	53.31	54.26	46.78	47.53
Fine-tuning	*White-box*	*Flo2*		*90.32*	*91.02*	*93.07*	*93.42*	*88.19*	*88.49*	N/A	N/A
LLM-wrapper	Black-box	Flo2	Mixtral	78.96 ↑11.1	77.69 ↑11.5	68.85↑12.9	68.21↑11.0	57.58↑4.3	58.26↑4.0	61.65 ↑14.9	65.14 ↑17.6
LLM-wrapper	Black-box	Flo2	Llama3	78.76 ↑10.9	78.03 ↑11.9	71.74↑15.8	71.91↑14.7	62.63↑9.3	62.73↑8.5	61.74 ↑15.0	65.84 ↑18.3

(‘(T)’, ‘(B)’, ‘GD’ and ‘Flo2’ stand for ‘SwinT’, ‘SwinB’, ‘Grounding DINO’ and ‘Florence-2’ respectively)

- **LLM-wrapper is model agnostic:** it boosts performances for all combinations of VLMs / LLMs
- **LLM-wrapper is not meant to compete with white-box fine-tuning**, but **it can reach on par results** in some cases (e.g. Grounding DINO on RefCOCOg despite a smaller backbone)

[1] Jiang et al. Mixtral of Experts. CoRR 2401.04088, 2024.
[2] Grattafiori et al. The Llama 3 Herd of Models. CoRR 2407.21783, 2024.

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Results

REC P@1↑ of LLM-wrapper, applied to VLMs already fine-tuned for REC

Adaptation	VLM	LLM	RefCOCOg val-umd	RefCOCOg test-umd
White-box FT	GD (B)		N/A†	66.30†
White-box FT + CRG (Wan et al., 2024)	GD (B)		N/A†	69.60† ↑3.30
White-box FT	GD (B)		78.51	77.99
White-box FT + LLM-wrapper	GD (B)	Mixtral	82.31 ↑3.80	82.15 ↑4.16
White-box FT + LLM-wrapper	GD (B)	Llama3	82.76 ↑4.25	82.61 ↑4.62
White-box FT	Flo2 FT		90.32	91.02
White-box FT + LLM-wrapper	Flo2 FT	Mixtral	90.40 ↑0.08	90.92 ↓0.10
White-box FT + LLM-wrapper	Flo2 FT	Llama3	90.50 ↑0.18	91.03 ↓0.01

(† marks results directly taken from Contrastive Region Guidance (CRG) [1] paper)

- LLM-wrapper is **compatible with other adaptation methods** (e.g. white-box REC fine-tuning)
- **It is a competitive black-box approach**

[1] Wan et al. Contrastive region guidance: Improving grounding in vision-language models without training. ECCV 2024.

ICLR 2025

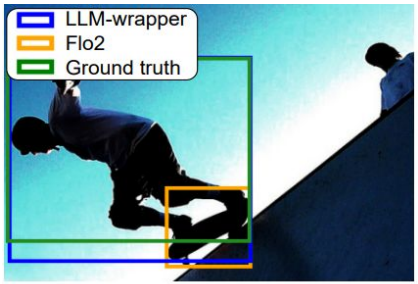
LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Results

LLM-wrapper enables:

- **target identification** (Fig. a & d)
- **spatial understanding** (Fig. b & e)
- **relational reasoning** (Fig. a & c)
- **It avoids choosing more visible objects** when ground truth is small (Fig. d & f)



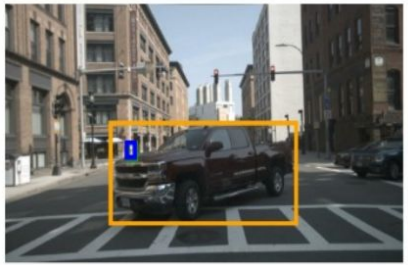
(a) *"Person on the skateboard"*

(b) *"The tie at the second from the left"*

(c) *"Green plant behind a table visible behind a lady's head"*

(d) *"She said something about a big sign on a fence. Maybe that is her ! Pull over here by this person and we will find out"*

(e) *"Try to get in front of the car that passed us on the left. He is driving like a madman"*



(f) *"My friend said she would be standing on the corner waiting for me, I think that might be her, will you stop there ?"*

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

ICLR 2025

LLM-wrapper enables ensembling

Adaptation	VLM	P@1 - val ↑	P@1 - test ↑
∅ (zero-shot VLM)	GDrec	67.61	68.37
∅ (zero-shot VLM)	Flo2	67.91	66.16
LLM-wrapper	GDrec	78.25	78.01
LLM-wrapper	Flo2	78.76	78.03
LLM-wrapper	Flo2 + GDrec	81.25	80.13

Llama 3, in P@1↑ on RefCOCOg. (Comparable findings for Mixtral).

LLM-wrapper can

- reason on multiple sources
- leverage the strengths of different VLMs

Query: *"A bottle of wine between the vegetables"*

(a) Ground truth (b) All candidates of Flo2 (c) All candidates of GDrec

(d) Final pred. of LLM-wrapper (e) Final pred. of Flo2 (f) Final pred. of GDrec

ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

LLM-wrapper enables VLM transfer

P@1↑ of LLM-wrapper, when using different VLMs outputs during fine-tuning and inference

Adaptation	VLM (fine-tuning)	VLM (inference)	P@1 - val ↑ (subset 300)	P@1 - val ↑ (full)	P@1 - test ↑ (full)
∅ (zero-shot VLM)	∅	GDrec	66.00†	67.61	68.37
LLM-wrapper	Flo2	GDrec	74.00† ↑8.0	73.90 ↑6.3	73.45 ↑5.1
∅ (zero-shot VLM)	∅	Flo2	71.67†	67.91	66.16
LLM-wrapper	GDrec	Flo2	75.33† ↑3.7	73.86 ↑6.0	73.03 ↑6.9
∅ (zero-shot VLM)	∅	GD-1.5	47.67†	—	—
LLM-wrapper	GDrec	GD-1.5	76.67† ↑29.0	—	—

(† marks scores obtained on a subset of 300 samples from RefCOCOg val)

LLM-wrapper can transfer from one VLM to another

- White-box fine-tuning is not possible in some cases. E.g. **Grounding-DINO 1.5** (GD-1.5) is **behind API**.
- Using LLM-wrapper on Grounding-DINO 1.5 requires many API calls (**≈ $1,600 to infer on RefCOCOg train**)
- One can **fine-tune LLM-wrapper on Grounding DINO (for "free") and use it on Grounding-DINO 1.5**

ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

LLM-wrapper enables dataset transfer

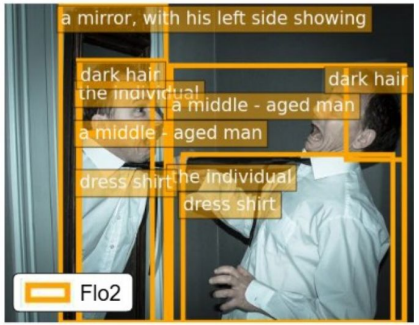
LLM-wrapper performance on zero-shot dataset transfer

Adaptation	Fine-tuning Data	Inference Data	P@1 - val ↑	P@1 - test ↑
∅ (zero-shot VLM)	—	HC-RefLoCo	48.04	47.39
White-box FT	*RefCOCO/+/g*†	*HC-RefLoCo*	*56.75* ↑**8.7**	*55.62* ↑**8.2**
`LLM-wrapper`	RefCOCOg	HC-RefLoCo	66.93 ↑**18.9**	66.45 ↑**19.1**

(HC-RefLoCo has 10x longer queries than RefCOCOg)

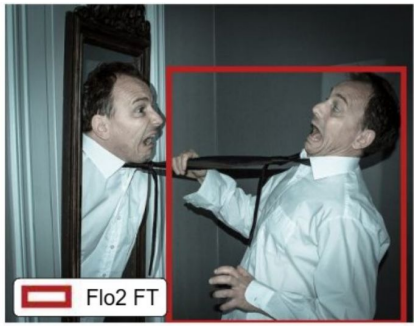
- LLM-wrapper displays **strong generalization across datasets** with different properties
- **White-box fine-tuning's boost** over zero-shot Florence 2 **is less than half of LLM-wrapper's**

(a) Ground truth



(b) All 10 candidates from `Flo2`

(c) Final pred. of `LLM-wrapper`



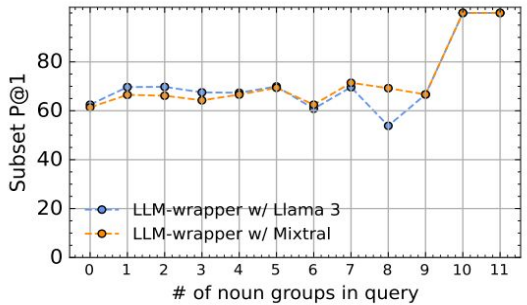
(d) Final pred. of fine-tuned `Flo2`

Query: "The individual is a middle-aged man with short, dark hair, appearing startled or comically alarmed. He is wearing a pale dress shirt and is positioned as if emerging from a mirror, with his left side showing."

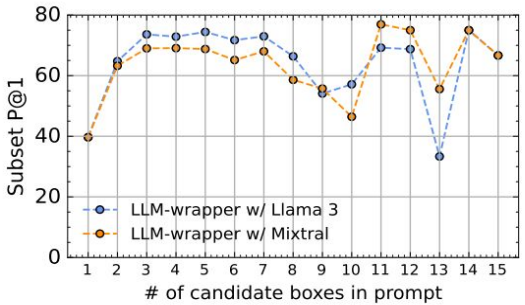
LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

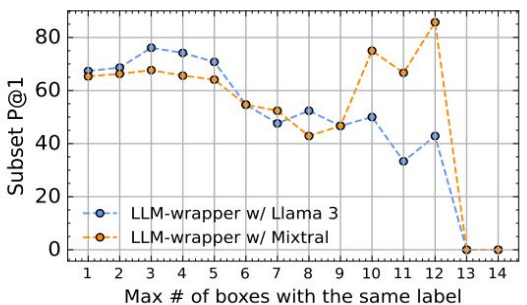
Ablation



(a) Query complexity.



(b) Number of candidate boxes.



(c) Box label redundancy.

Figure 6: **Study of `LLM-wrapper`'s sensitivity to input complexity.**

- LLM-wrapper is robust to, and even benefits from, increasing textual complexity
- LLM-wrapper is robust to an increasing number of box candidates in the prompt
- Decreasing performances are observed as more boxes share a same label

ICLR 2025

LLM-wrapper: Black-Box Semantic-Aware Adaptation of VLMs for Referring Expression Comprehension

Amaia Cardiel[1,2], Eloi Zablocki[1], Elias Ramzi[1], Oriane Siméoni[1], Matthieu Cord[1,3]
[1]Valeo.ai [2]Université Grenoble Alpes [3]Sorbonne Université

Takeaway

- **Improving semantic understanding** is key to the REC task
- It enables **black-box adaptation** of **any 0 shot or fine-tuned VLMs** on REC
- **LLMs can learn highly transferable spatial and semantic knowledge**

Project Website

Thank you for your attention !