Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

The devil is in the fine-grained details

Fabrizio Falchi
CNR

23/9/2024 Napoli

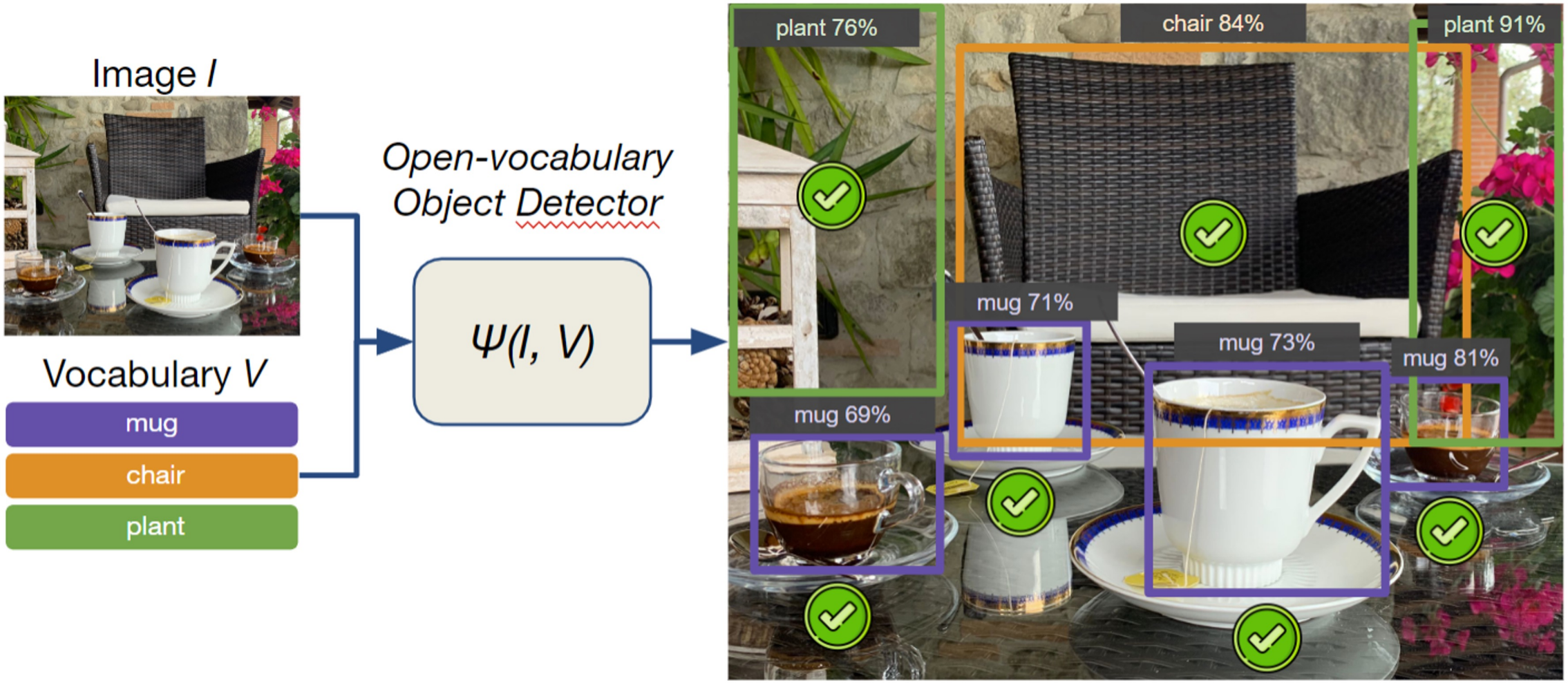
"There's really a philosophical point that you could learn a very good model from language alone, but it's much easier to learn it from a multimodal system."

Geoffrey Hinton

at AI Summit, May 15 2024 in Stockholm

https://www.youtube.com/watch?v=n4IQOBka8bc

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research



Bianchi et al., *The devil is in the fine-grained details. Evaluating open-vocabulary object detectors for fine-grained understanding.* CVPR 2024

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Image *I*

Vocabulary *V*

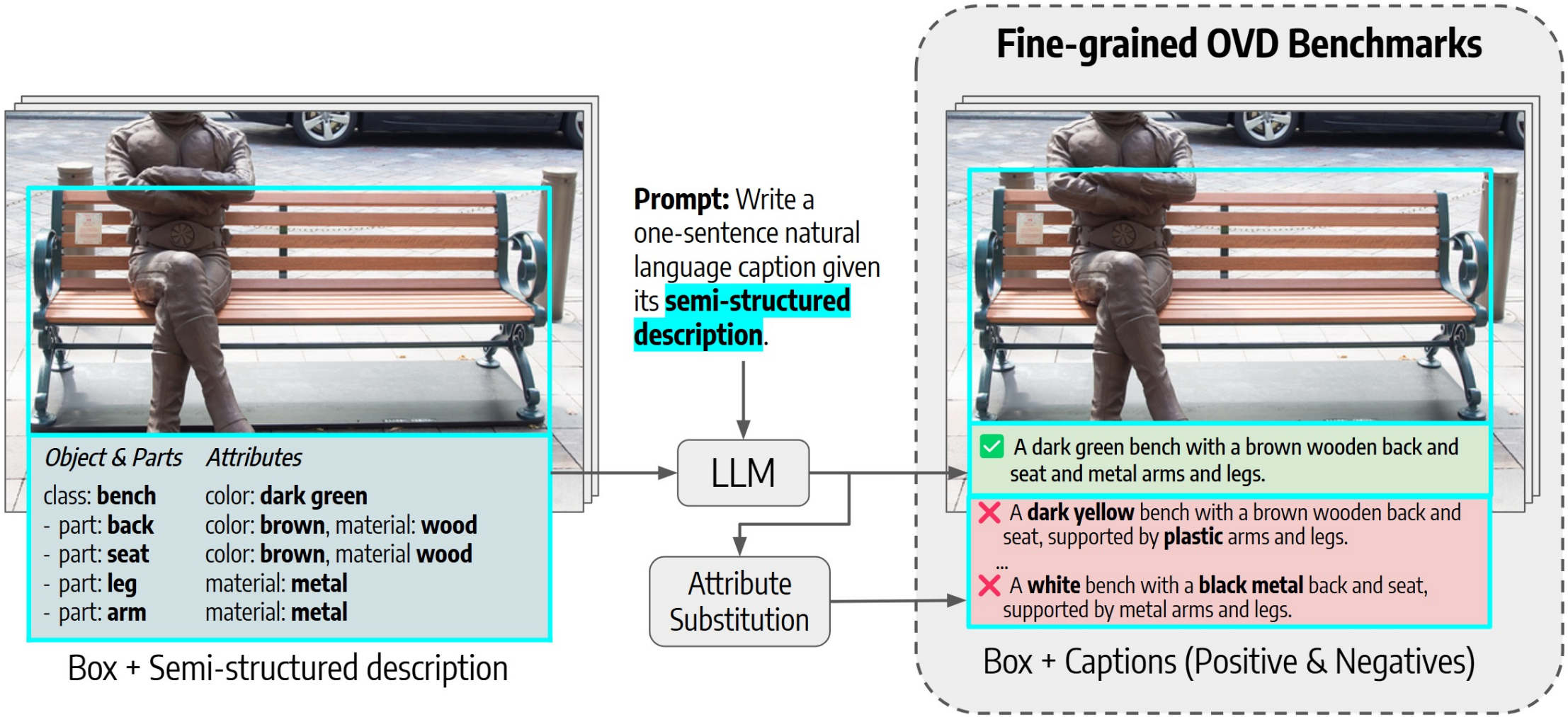
a mug with blue rim

a plant with red flowers

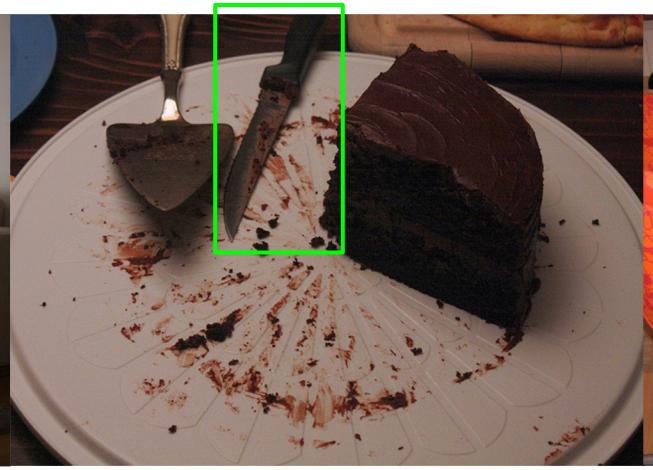
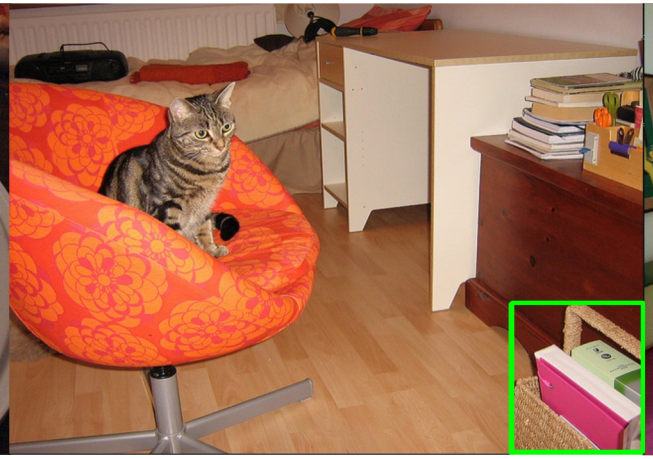
Open-vocabulary Object Detector

$\Psi(I, V)$

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Hard	Medium	Easy	Trivial
			
✅ A lamp with a white plastic shade and a grey metal pipe	✅ A knife with a black plastic handle and a dark grey metal blade	✅ A brown woven rattan basket	✅ A brown wooden chair
❌ A lamp with a white **velvet** shade and a grey metal pipe	❌ A knife with a **grey stone** handle and a dark grey metal blade	❌ A **light green perforated fabric** basket	❌ **A red cup with a pink plastic rim**
❌ A lamp with a white plastic shade and a **dark pink** metal pipe	❌ A knife with a **light pink** plastic handle and a dark **light yellow** metal blade	❌ A **black dotted leather** basket	❌ **A pink dog with a black and dotted ear**

Bianchi et al., *The devil is in the fine-grained details. Evaluating open-vocabulary object detectors for fine-grained understanding.* CVPR 2024

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Color

✅ A blue hat

❌ A **orange** hat

❌ A **yellow** hat

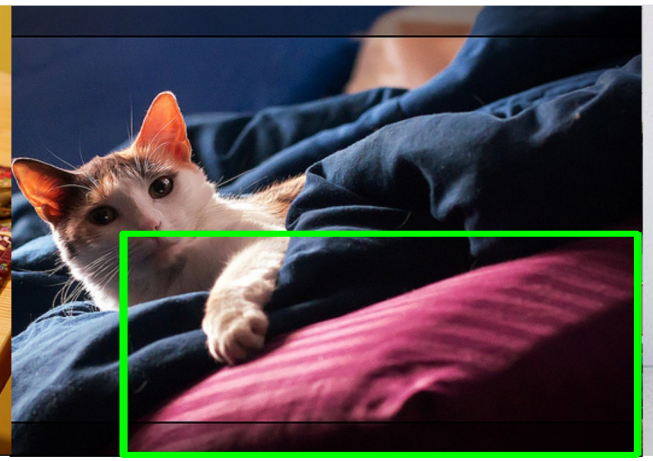
Material

✅ A red plastic plate

❌ A red **metal** plate

❌ A red **ceramic** plate

Pattern

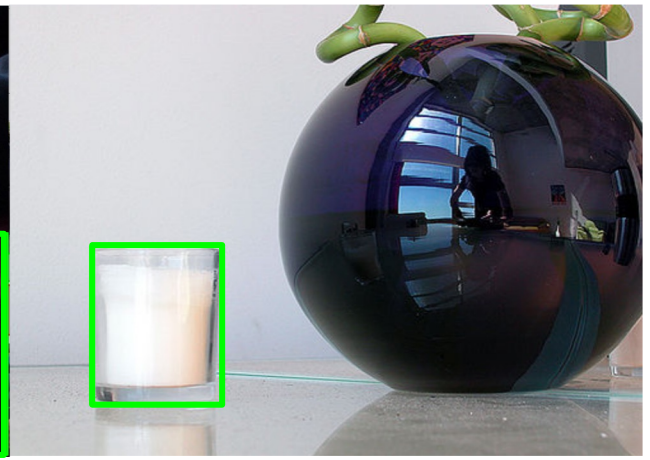


✅ A dark pink striped pillow.

❌ A dark pink **floral** pillow.

❌ A dark pink **dotted** pillow.

Transparency



✅ A transparent glass

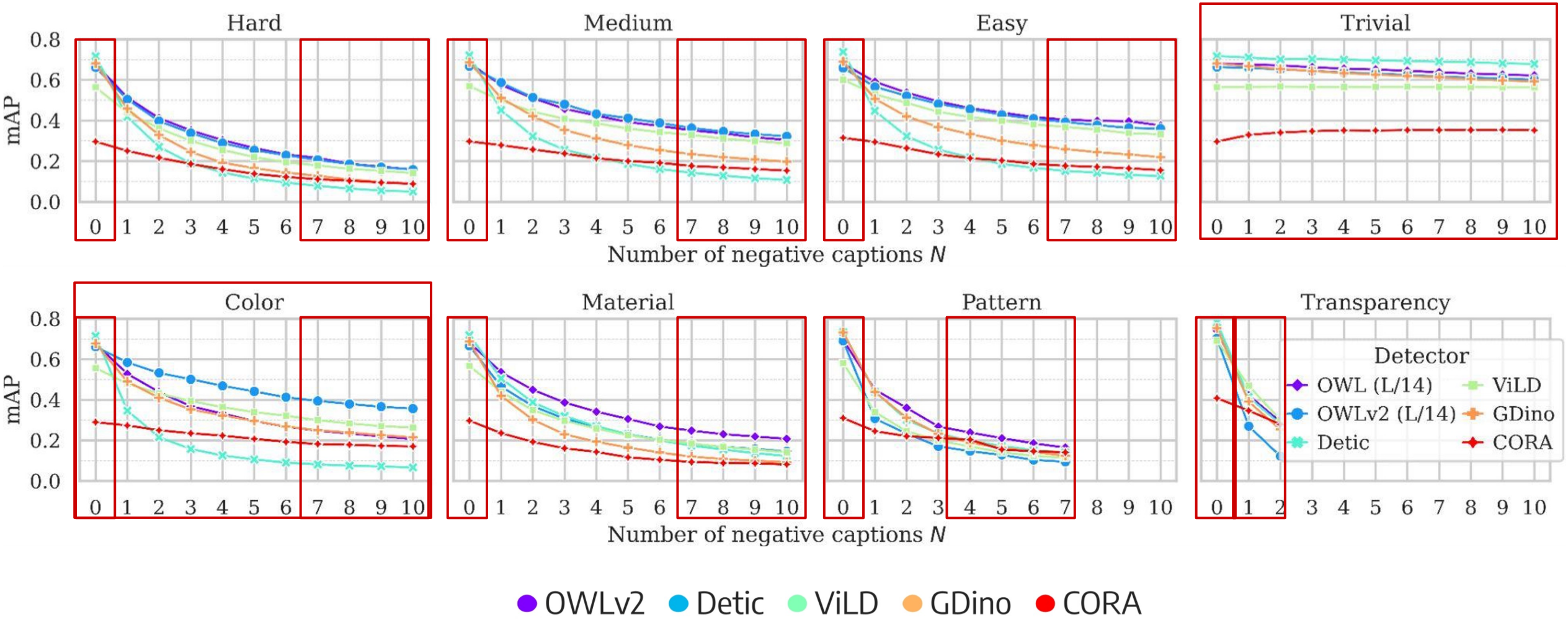
❌ A **translucent** glass

❌ A **opaque** glass

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

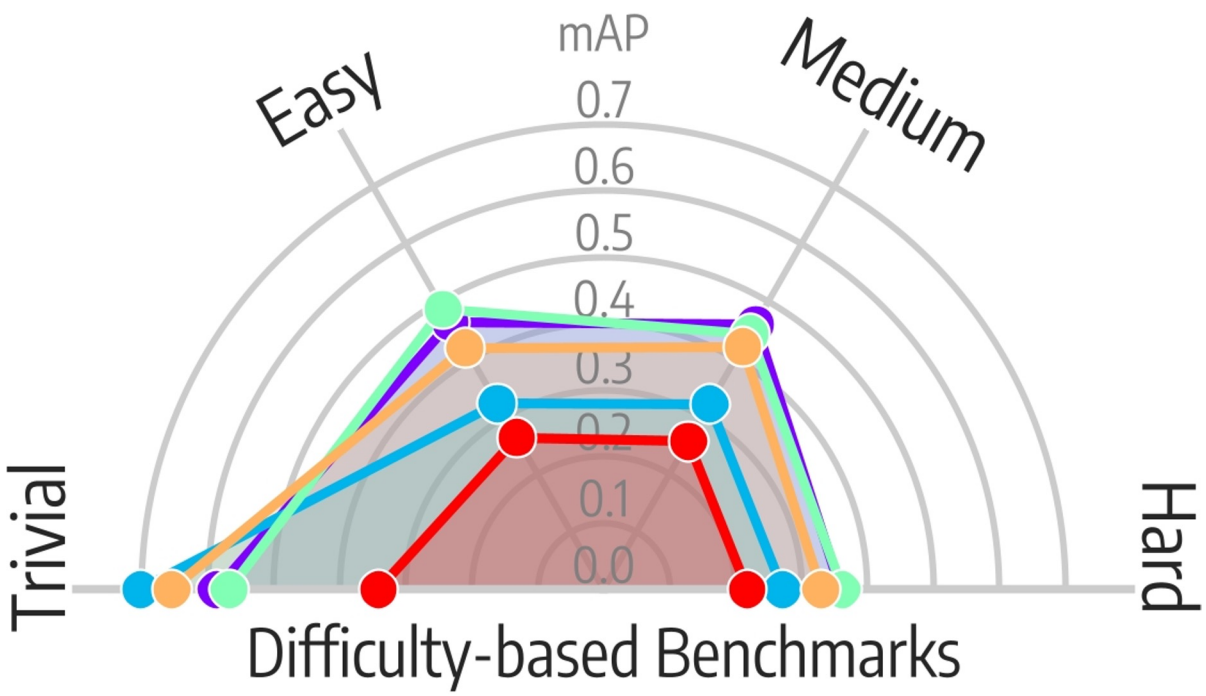
Name	Negative Set Strategy	Imgs	Objs	Obj/Img	✅ Caps	✅ /Img	❌/✅	Objs/✅
Hard	Random attribute subst. (x1)	1707	3545	2.1	2349	1.4	9.9	1.5
Normal	Random attribute subst. (x2)	1537	2968	1.9	2034	1.3	10.0	1.5
Easy	Random attribute subst. (x3)	853	1299	1.5	971	1.1	10.0	1.3
Trivial	Random captions	1707	3545	2.1	2349	1.4	9.9	1.5
Color	Color attribute subst.	1599	3119	2.0	2126	1.3	10.0	1.5
Material	Material attribute subst.	1577	3193	2.0	2128	1.3	10.0	1.5
Pattern	Pattern attribute subst.	321	467	1.5	337	1.0	7.4	1.4
Transparency	Transparency attribute subst.	230	409	1.8	238	1.0	2.2	1.7

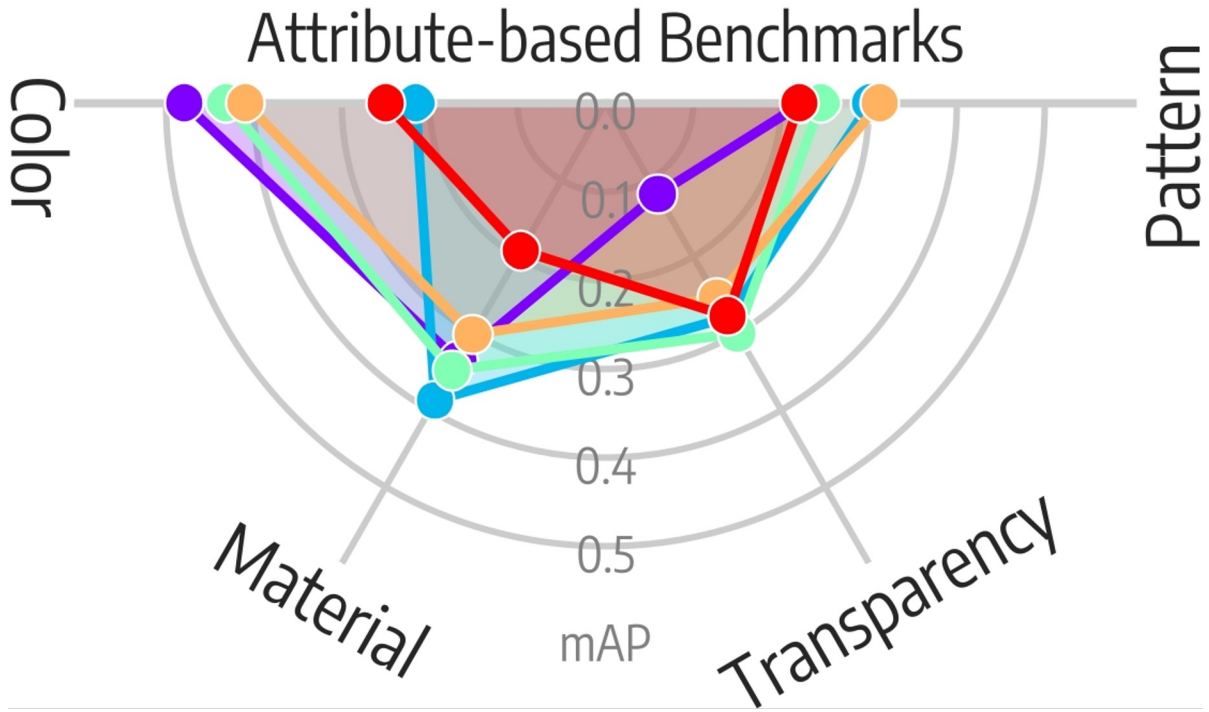
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research



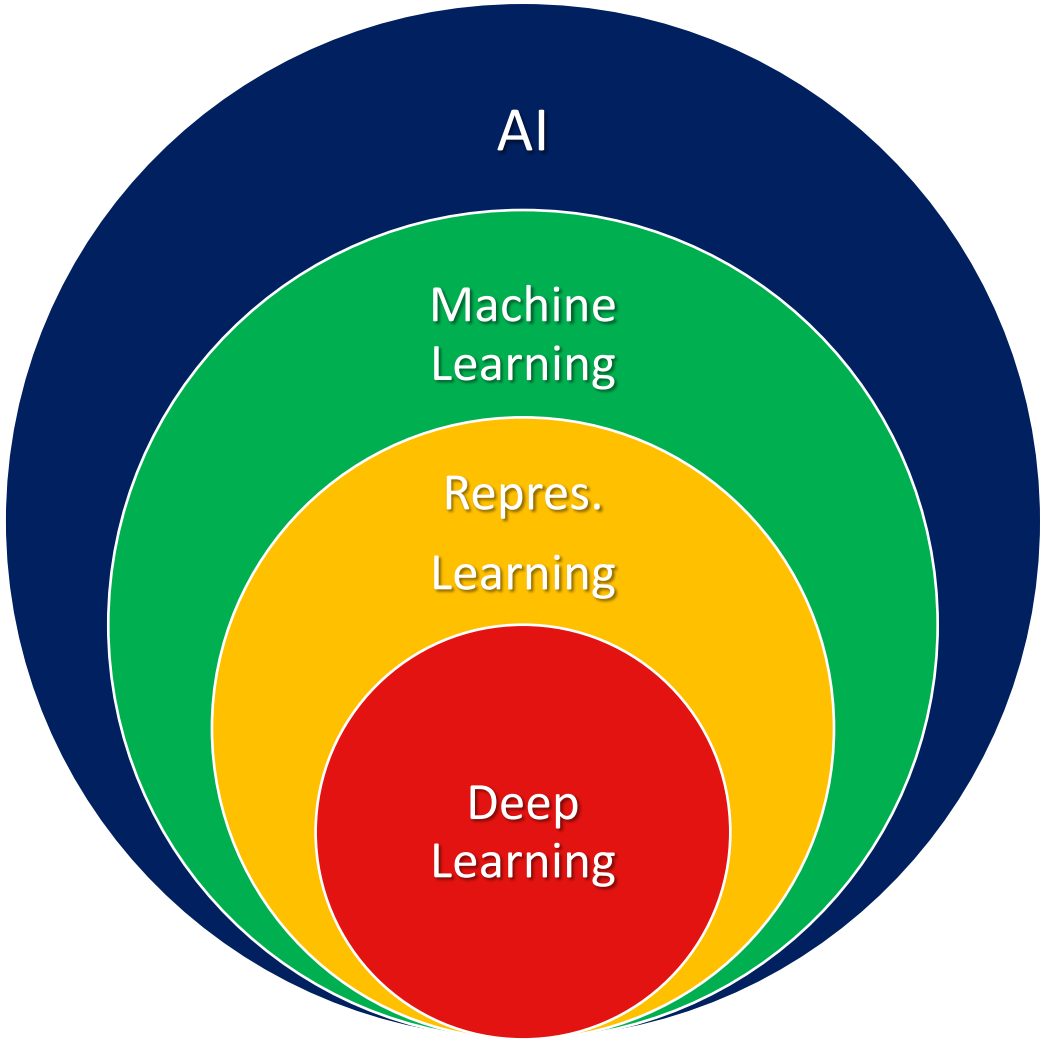
Bianchi et al., *The devil is in the fine-grained details. Evaluating open-vocabulary object detectors for fine-grained understanding.* CVPR 2024

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

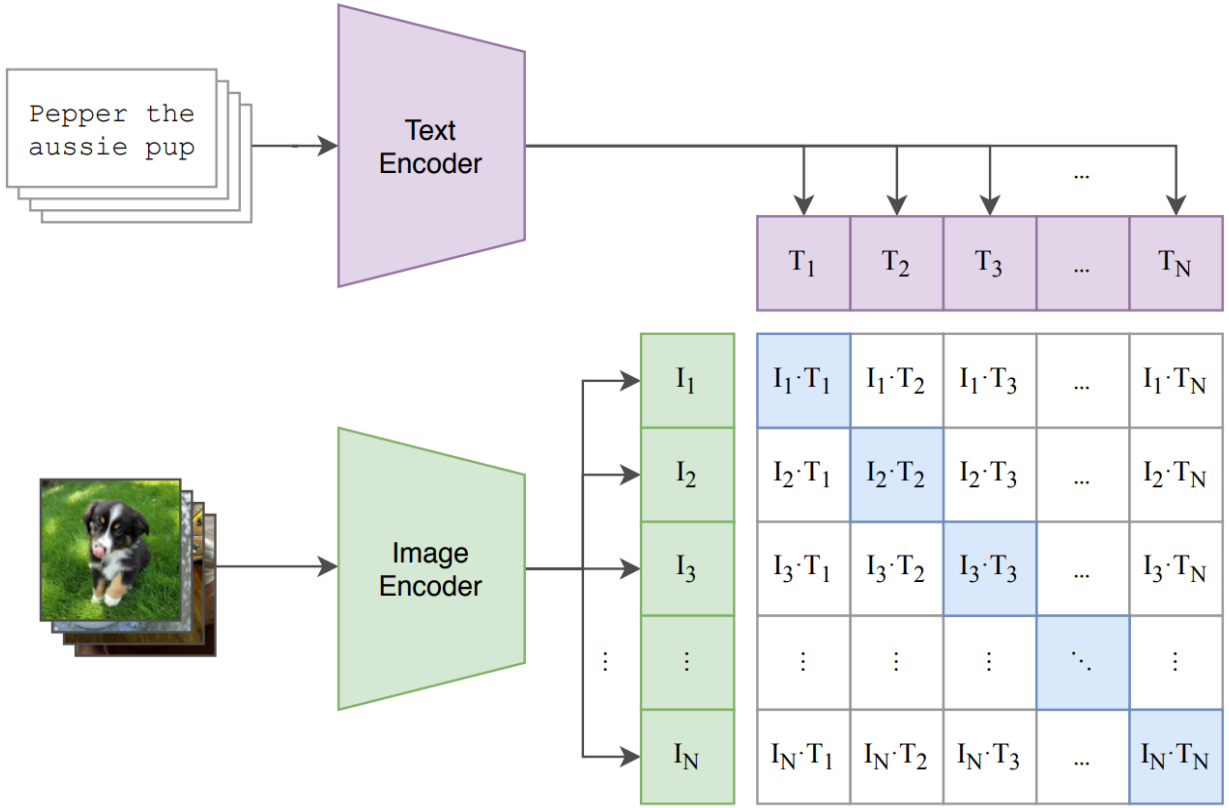




OWLv2 Detic ViLD GDino CORA



CLIP: Contrastive Language-Image Pre-Training



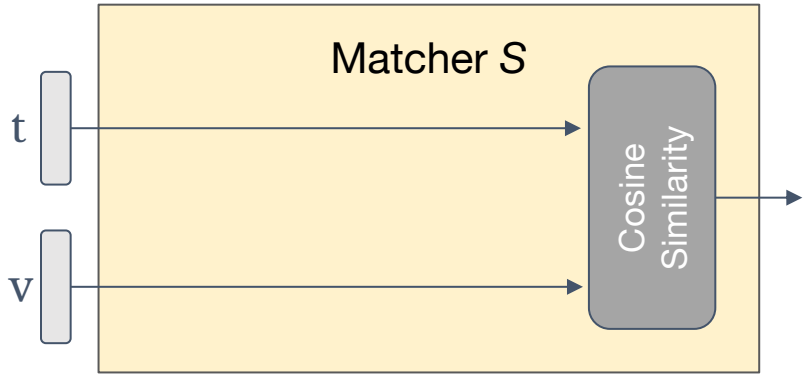
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Fine-grained understanding (lower is better)

Standard coarse-grained understanding (higher is better)

	FG-OVD	COCO Retrieval					
		I→T			T→I		
	Mean Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
CLIP B/16	2.98	41.5	65.9	76.2	22.6	44.1	54.9

Baseline, no training



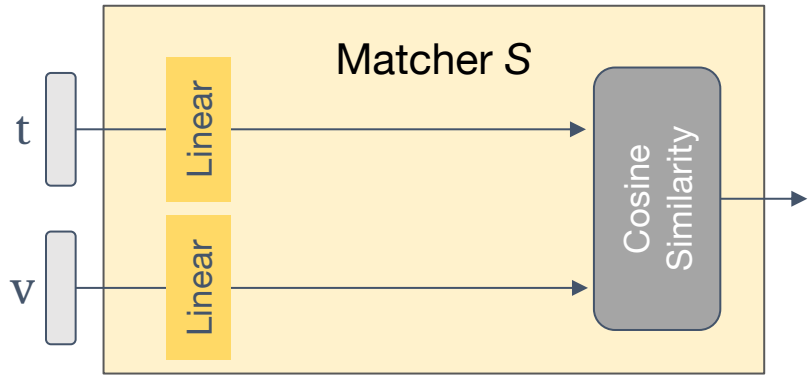
Answering Q2: ablating matching

Fine-grained understanding (lower is better)

Standard coarse-grained understanding (higher is better)

	FG-OVD	COCO Retrieval					
		I→T			T→I		
	Mean Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
CLIP B/16	2.98	41.5	65.9	76.2	22.6	44.1	54.9
Linear (both)	3.78	**48.0**	**76.6**	**85.4**	**37.2**	**65.6**	**76.6**
+FG-OVD	1.46 (-2.32)	37.1 (-10.9)	66.8 (-9.8)	78.4 (-7.0)	35.6 (-1.6)	63.9 (-1.7)	75.0 (-1.6)

Warm-up new parameters on COCO (first line)
Fine-tune them on FG-OVD (second line)

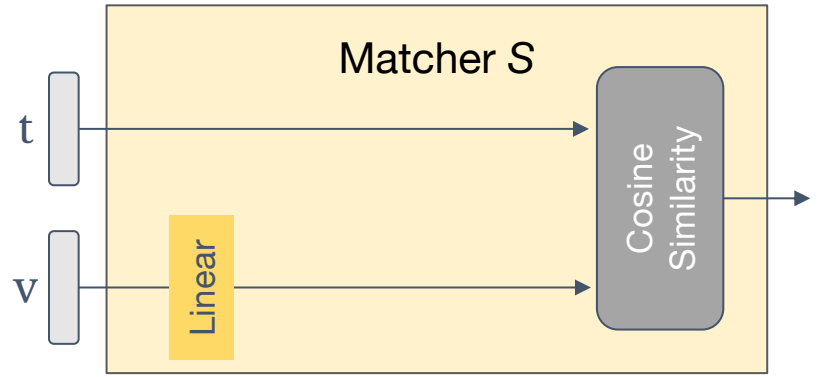


Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Fine-grained understanding (lower is better)

Standard coarse-grained understanding (higher is better)

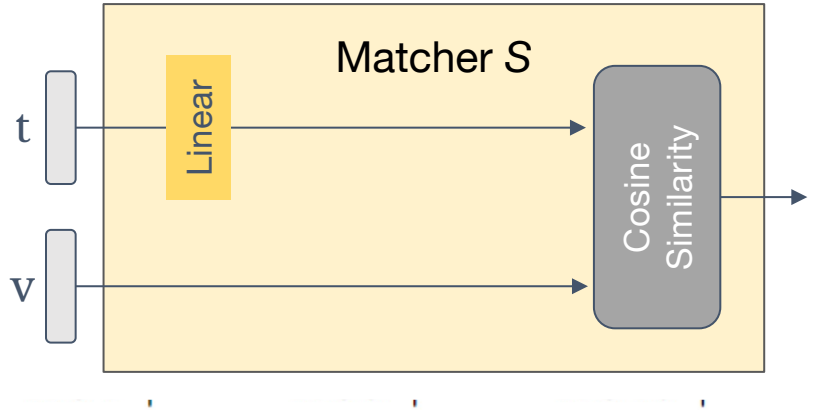
	FG-OVD	COCO Retrieval					
		I→T			T→I		
	Mean Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
CLIP B/16	2.98	41.5	65.9	76.2	22.6	44.1	54.9
Linear (both)	3.78	**48.0**	**76.6**	**85.4**	**37.2**	**65.6**	**76.6**
+FG-OVD	1.46 (-2.32)	37.1 (-10.9)	66.8 (-9.8)	78.4 (-7.0)	35.6 (-1.6)	63.9 (-1.7)	75.0 (-1.6)
Linear (visual only)	3.53	45.8	74.2	83.7	35.4	64.2	75.3
+FG-OVD	1.54 (-1.99)	39.4 (-6.4)	69.5 (-4.7)	79.8 (-3.9)	34.3 (-1.1)	62.9 (-1.3)	74.1 (-1.2)



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Fine-grained understanding (lower is better)

Standard coarse-grained understanding (higher is better)

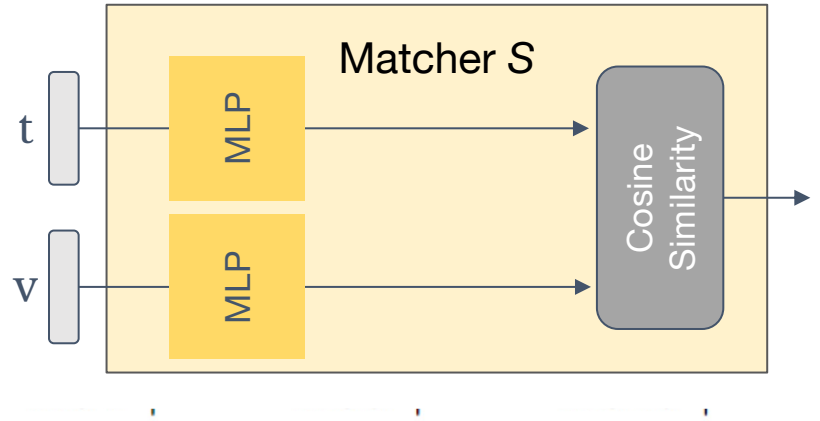


	FG-OVD	COCO Re				
		I→T				
	Mean Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑		
CLIP B/16	2.98	41.5	65.9	76.2	22.6	44.1	54.9
Linear (both)	3.78	**48.0**	**76.6**	**85.4**	**37.2**	**65.6**	**76.6**
+FG-OVD	1.46 (-2.32)	37.1 (-10.9)	66.8 (-9.8)	78.4 (-7.0)	35.6 (-1.6)	63.9 (-1.7)	75.0 (-1.6)
Linear (visual only)	3.53	45.8	74.2	83.7	35.4	64.2	75.3
+FG-OVD	1.54 (-1.99)	39.4 (-6.4)	69.5 (-4.7)	79.8 (-3.9)	34.3 (-1.1)	62.9 (-1.3)	74.1 (-1.2)
Linear (text only)	3.48	47.3	75.1	84.9	36.0	64.3	75.6
+FG-OVD	1.57 (-1.91)	41.1 (-6.2)	70.4 (-4.7)	80.6 (-4.3)	34.7 (-1.3)	63.2 (-1.1)	74.6 (-1.0)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Fine-grained understanding (lower is better)

Standard coarse-grained understanding (higher is better)

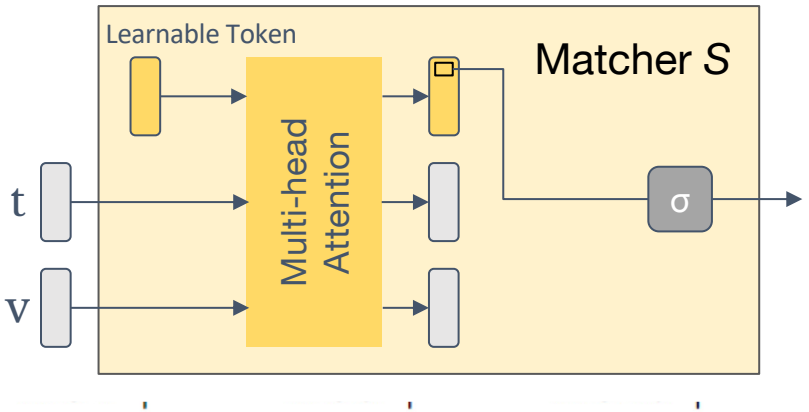


	FG-OVD	COCO Re				
		I→T				
	Mean Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑		
CLIP B/16	2.98	41.5	65.9	76.2	22.6	44.1	54.9
Linear (both)	3.78	**48.0**	**76.6**	**85.4**	**37.2**	**65.6**	**76.6**
+FG-OVD	1.46 (-2.32)	37.1 (-10.9)	66.8 (-9.8)	78.4 (-7.0)	35.6 (-1.6)	63.9 (-1.7)	75.0 (-1.6)
Linear (visual only)	3.53	45.8	74.2	83.7	35.4	64.2	75.3
+FG-OVD	1.54 (-1.99)	39.4 (-6.4)	69.5 (-4.7)	79.8 (-3.9)	34.3 (-1.1)	62.9 (-1.3)	74.1 (-1.2)
Linear (text only)	3.48	47.3	75.1	84.9	36.0	64.3	75.6
+FG-OVD	1.57 (-1.91)	41.1 (-6.2)	70.4 (-4.7)	80.6 (-4.3)	34.7 (-1.3)	63.2 (-1.1)	74.6 (-1.0)
MLP	3.49	45.9	75.5	84.6	36.5	64.9	76.2
+FG-OVD	**1.43 (-2.06)**	31.9 (-14.0)	60.2 (-15.3)	72.7 (-11.9)	33.6 (-2.9)	62.0 (-2.9)	73.9 (-2.3)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Fine-grained understanding (lower is better)

Standard coarse-grained understanding (higher is better)



	FG-OVD	COCO Re				
		I→T				
	Mean Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑		
CLIP B/16	2.98	41.5	65.9	76.2	22.6	44.1	54.9
Linear (both)	3.78	**48.0**	**76.6**	**85.4**	**37.2**	**65.6**	**76.6**
+FG-OVD	1.46 (-2.32)	37.1 (-10.9)	66.8 (-9.8)	78.4 (-7.0)	35.6 (-1.6)	63.9 (-1.7)	75.0 (-1.6)
Linear (visual only)	3.53	45.8	74.2	83.7	35.4	64.2	75.3
+FG-OVD	1.54 (-1.99)	39.4 (-6.4)	69.5 (-4.7)	79.8 (-3.9)	34.3 (-1.1)	62.9 (-1.3)	74.1 (-1.2)
Linear (text only)	3.48	47.3	75.1	84.9	36.0	64.3	75.6
+FG-OVD	1.57 (-1.91)	41.1 (-6.2)	70.4 (-4.7)	80.6 (-4.3)	34.7 (-1.3)	63.2 (-1.1)	74.6 (-1.0)
MLP	3.49	45.9	75.5	84.6	36.5	64.9	76.2
+FG-OVD	**1.43 (-2.06)**	31.9 (-14.0)	60.2 (-15.3)	72.7 (-11.9)	33.6 (-2.9)	62.0 (-2.9)	73.9 (-2.3)
MHA	4.08	36.3	66.1	78.1	29.1	57.6	70.3
+FG-OVD	1.54 (-2.54)	22.3 (-14.0)	48.3 (-17.8)	61.2 (-16.9)	22.6 (-6.5)	49.2 (-8.4)	62.0 (-8.3)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Fabrizio Falchi
fabrizio.falchi@cnr.it

AIMH
ARTIFICIAL INTELLIGENCE FOR MEDIA AND HUMANITIES

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" Consiglio Nazionale delle Ricerche