CTRL-O: Language-Controllable Object-Centric Visual Representation Learning

Aniket Didolkar*, Andrii Zadaianchuk*, Rabiul Awal*, Maximilian Seitzer, Efstratios Gavves, Aishwarya Agrawal
*denotes equal contribution

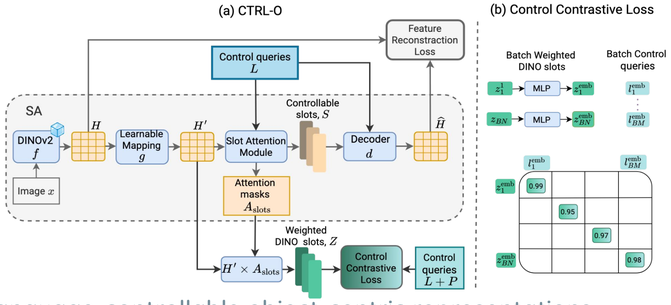
Mila Université de Montréal
University of Amsterdam

Object-Centric Learning

Pros:
- Image decomposition into **a set of slots** where each slots represents a unique object.
- Works well on real-world images.

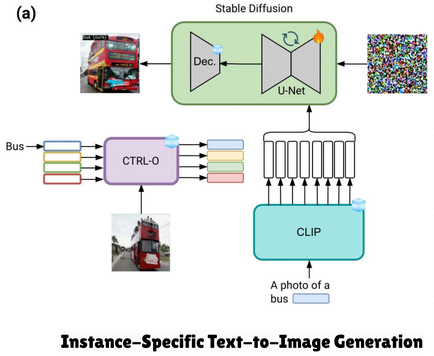
Cons:
- Cannot be guided by user queries.
- Always recovers a fixed decomposition of a scene.

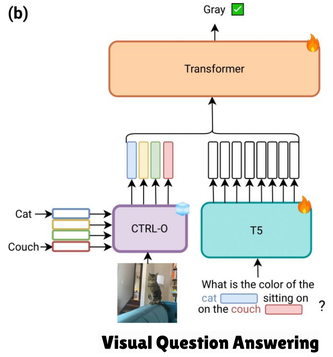
CTRL-O



CTRL-O transforms visual features into language-controllable object-centric representations.
Enables precise binding between language descriptions and specific objects in complex scenes.

Downstream Applications: Image Generation & VQA





CTRL-O allows us to extract features for specific (instances of) objects and either generate new pictures of those instances of objects or answer questions about those objects.

Coupling: Extract object specific visual features from CTRL-O and use them alongside the corresponding text in the question for richer image-text interactions.

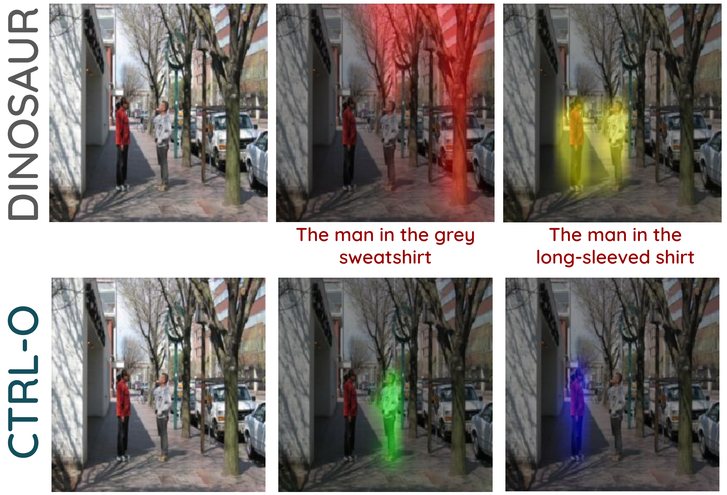
Object Discovery and Visual Grounding

Slot Init.	GT Masks	CL	DC	Binding Hits	FG-ARI	mBO
✓	✓	✗	✗	**71.2**	**69.8**	**35.4**
✓	✗	✗	✗	8.1	34.52	22.42
✓	✗	✗	✓	10.11	43.83	25.76
✓	✗	✓	✗	56.3	44.8	27.3
✓	✗	✓	✓	61.3	47.5	27.2

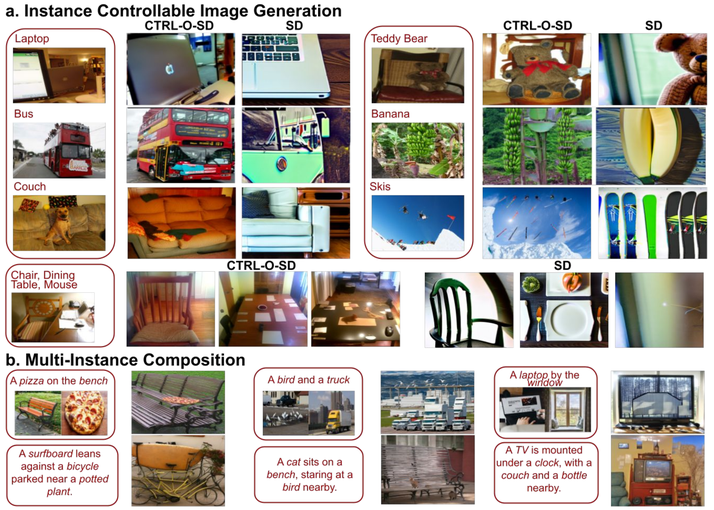
Ablation over various CTRL-O components for achieving strong visual grounding.

Approach	Model	FG-ARI	mBO
Unsup.	DINOSAUR (MLP Dec.) [38]	40.5	27.7
	DINOSAUR (TF. Dec.) [38]	34.1	31.6
	Stable-LSD [19]	35.0	30.4
	SlotDiffusion [43]	37.3	31.4
Weak Sup.	Stable-LSD (Bbox Supervision) [39]	-	30.3
	CTRL-O (Trained on COCO)	47.5	27.2

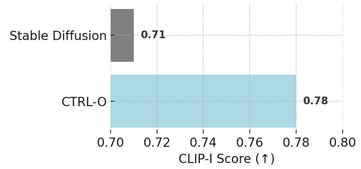
CTRL-O demonstrates strong object discovery.



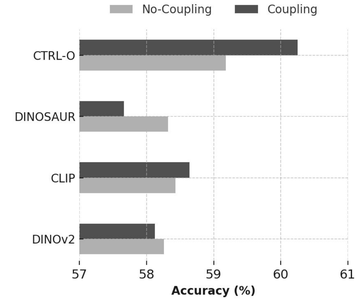
With CTRL-O, we can control the granularity at which objects are discovered in the image.



a. CTRL-O enables generating new images for specific instances of objects in source images.
b. Composes multiple extracted instances from different images into a coherent scene.



CTRL-O generates images that better match the target instance.



CTRL-O improves VQA performance by enabling coupling.