Causal Graphical Models for Vision Language Compositional Understanding

Fiorenzo Parascandolo Nicholas Moratelli Enver Sangineto Lorenzo Baraldi Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy

Compositionality in Vision Language Models

- Image-to-text retrieval between a positive and one (or more) negative captions
- The candidate captions contain all the same words (in a different order) or differ by a few words
- CLIP models performs poorly in compositional tasks [1]
- Generative image captioning approaches could solve this issue [2]

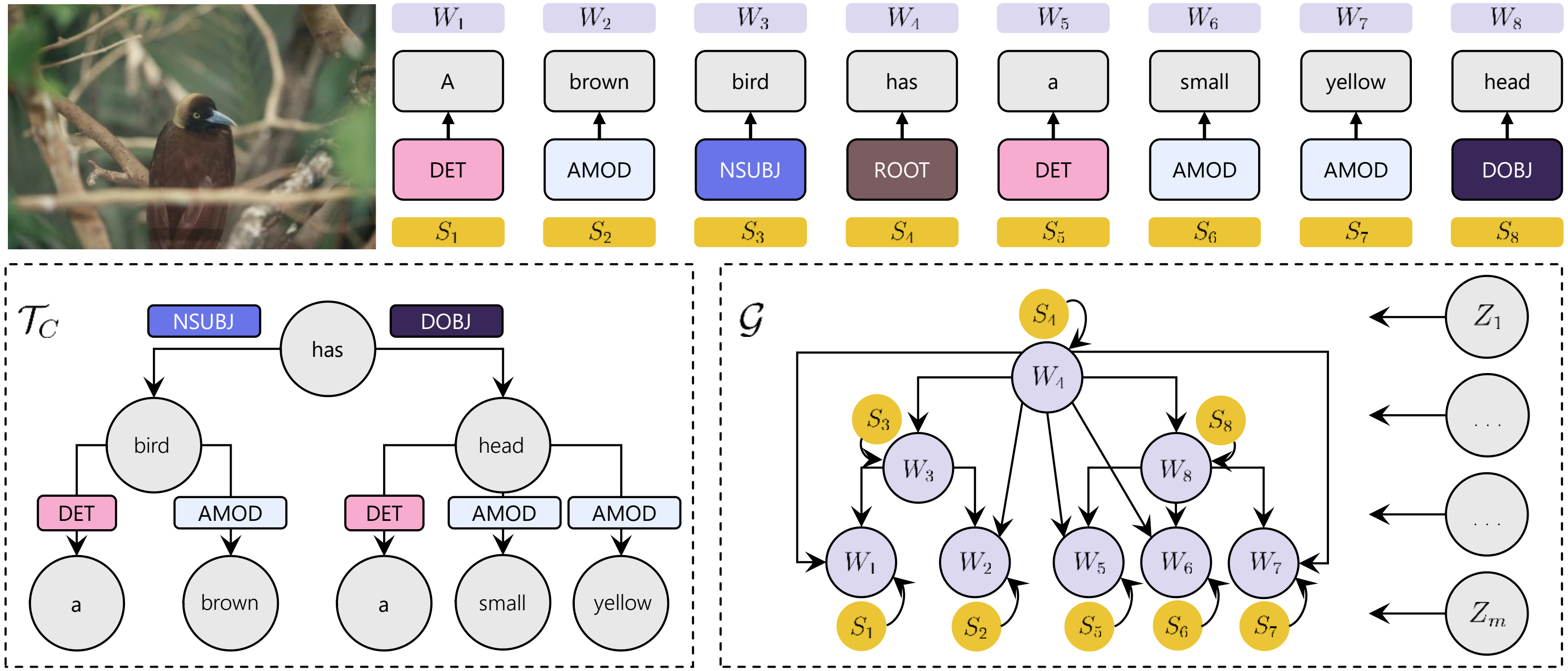
A brown bird has a small yellow <u>head</u>.

A brown bird has a small yellow <u>beak</u>.

[1] Mert Yuksekgonul et al. When and why vision-language models behave like bags-of-words, and what to do about it? ICLR 2023
[2] Michael Tschannen et al. Image captioners are scalable vision learners too. NeurIPS 2023

Motivations

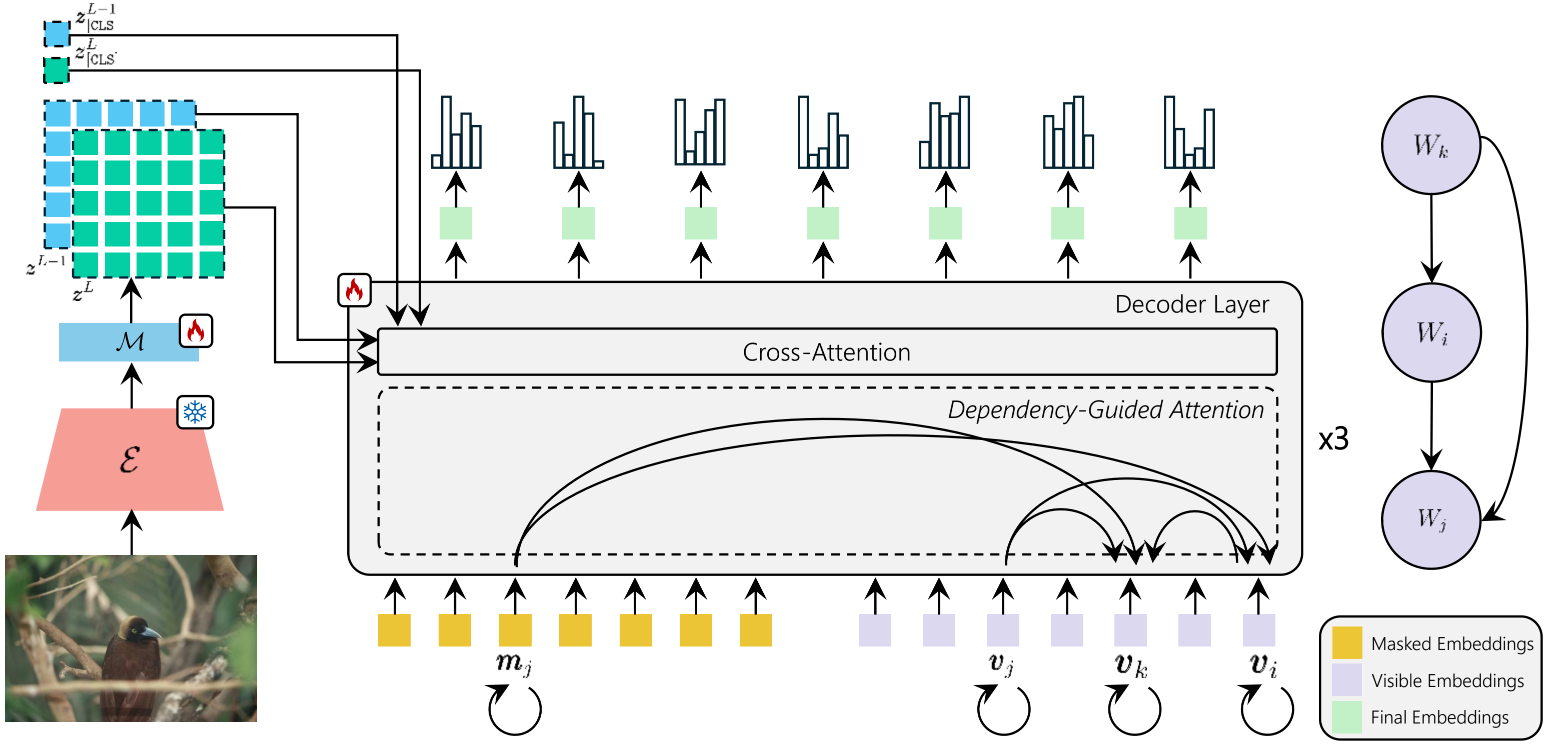
- Standard image captioning could be ambiguous: the model must predict "brown" before knowing that this adjective refers to "bird"
- A dependency tree shows how words in a sentence depend on one another
- A causal graphical model shows how one variable cause another
- We can see dependency relations as cause-and-effect links
- **We propose to use dependency relations between words to determine the order of token prediction**



Causally-Ordered Generative Training (COGT)

We call our approach **Causally Ordered Generative Training (COGT)**:

1. **Build Dependency Tree** using a dependency parser to determine the order of token prediction and syntactic label for each token
2. **Build Causal Graphical Model** to connect each word to its syntactic type, ancestors in the dependency tree, and all visual features
3. **Train a decoder** maximizing: $P(W_1,\dots,W_n|Z_1,\dots,Z_m) = \sum_{j=1}^{n} \log(P(W_j|\mathbf{PA}(W_j))), \quad \mathbf{PA}(W_j) = \{W_{i_1},\dots,W_{i_k},S_j,Z_1,\dots,Z_m\}$



Ablations

- **Five compositional benchmarks**: ARO, SugarCrepe, VL-CheckList and ColorSwap and an additional benchmark FG-OVD which we adapt for compositional tasks
- **Training Set**: COCO
- **Dependency Parser**: Deep Biaffine + RoBERTa [1] achieves the best performance, consistent with top Penn Tree Bank rankings
- **Mask Tokens**: Category-specific masked tokens yield a +2.69 accuracy
- **Layers**: Using both the final and penultimate visual features from the frozen visual backbone yield a +4.75 accuracy compared to using only the last layer
- **Visual Backbone**: CLIP, ViT B/32 [2]

			ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
Parser	**Mask-Specific**	**Layers**	**Relation**	**Attribute**	**Avg**	**Add**	**Replace**	**Swap**	**Avg**	**Attribute**	**Object**	**Relation**	**Avg**	**ITT**	**Avg**	
CRFPar	✓	2	85.68	88.34	87.01	98.16	84.94	80.30	87.80	86.99	77.68	87.09	83.92	56.33	43.74	71.76
Deep Biaffine	✓	2	86.56	89.10	87.83	98.11	85.80	81.49	88.46	87.02	78.30	87.75	84.35	61.33	44.74	73.34
Deep Biaffine + RoBERTa	✗	2	84.75	86.16	85.46	98.86	84.37	80.25	87.82	83.79	78.24	90.84	84.29	58.00	46.99	72.51
Deep Biaffine + RoBERTa	✓	1	86.82	89.67	88.25	98.26	86.56	82.33	89.05	84.41	78.94	89.54	84.30	45.00	45.63	70.45
Deep Biaffine + RoBERTa	✓	2	87.56	90.26	88.91	98.26	87.10	83.14	89.50	86.07	78.91	89.37	84.78	61.33	51.48	75.20

[1] Timothy et al. Deep biaffine attention for neural dependency parsing. arXiv 2016.
[2] Alec Radford et al. Learning transferable visual models from natural language supervision. ICML 2021

COGT-CLIP

Model	ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
	Relation	Attribute	Avg	Add	Replace	Swap	Avg	Attribute	Object	Relation	Avg	ITT	Avg	
Zero-shot														
CLIP	59.00	62.00	60.50	85.58	80.76	70.83	79.05	67.93	82.83	64.19	71.65	35.67*	47.33	58.84
Training on COCO only														
CLIP Fine-Tuned	63.00	65.00	64.00	.	.	.	.	.	.	.	.	.	.	.
NegCLIP	81.00	71.00	76.00	87.29	85.36	75.30	82.65	72.24	87.00	71.39	76.87	35.67*	41.69	62.57
CE-CLIP	83.00	76.40	79.70	92.90	87.00	74.90	84.94	72.60	84.60	71.80	76.30	18.67	41.97	60.31
Structure-CLIP	85.10*	83.50*	84.30*	.	.	.	.	.	.	.	.	.	.	.
GNM	65.00	65.00	65.00	82.85	80.95	66.71	76.83	70.15	85.91	64.10	73.38	13.00	38.79	53.40
Plausible Adj. Neg	65.07	67.94	66.51	89.64	85.37	70.88	81.96	76.51	88.13	69.90	78.17	17.67	44.98	57.86
SDS-CLIP	55.00	66.00	60.50	.	.	.	.	.	.	.	.	.	.	.
COGT-CLIP	87.56	90.26	88.91	98.26	87.10*	83.14	89.50	86.07	78.91	89.37*	84.78	61.33	51.48	75.20
Training on datasets larger than COCO														
CE-CLIP+	83.60	77.10	80.35	94.40	**89.30**	78.00*	87.23*	76.70	86.30	74.70	79.23	.	.	.
IL-CLIP	.	.	.	73.80	73.00	62.90	69.90	.	.	.	.	.	.	.
syn-CyCLIP	69.00	63.65	66.33	.	.	.	.	68.06	.	65.73	.	.	.	.
CLIP-SPEC	73.70	66.40	70.05	.	.	.	.	.	.	.	.	.	.	
DAC-SAM	77.16	70.50	73.83	92.87	86.18	71.06	83.37	75.80	**88.50**	89.80	84.70*	16.33	48.36	61.31
DAC-LLM	81.28	73.91	77.60	95.83*	88.09	72.48	85.47	77.30*	87.30*	86.40	83.66	18.33	49.60*	62.93*
COGT-CLIP+	**90.67**	**96.01**	**93.34**	**98.42**	87.05	**84.21**	**89.89**	**90.71**	84.91	**92.33**	**89.31**	**81.66**	**69.96**	**84.83**

- Visual backbone: CLIP, ViT B/32 [1]
- COGT-CLIP is trained on COCO
- COGT-CLIP+ trained on CC3M + COCO + Visual Genome

[1] Alec Radford et al. Learning transferable visual models from natural language supervision. ICML 2021

COGT-XVLM

Model	ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
	Relation	Attribute	Avg	Add	Replace	Swap	Avg	Attribute	Object	Relation	Avg	ITT	Avg	
Zero-shot														
XVLM	73.40	86.80	80.10	.	.	.	.	75.10*	85.80	70.40	76.50	.	.	.
Training on COCO only														
CE-XVLM	73.90*	89.30*	81.60*	.	.	.	.	74.80	**86.90**	79.70*	78.60*	.	.	.
HardNeg-DiffusionITM	52.30	67.60	59.95	.	.	.	.	.	.	.	.	.	.	.
COGT-XVLM	87.64	92.30	89.97	**98.65***	**89.17**	84.37	90.73	85.87	80.49	88.74	85.03	69.67	50.12	77.10
Training on datasets larger than COCO														
COGT-XVLM+	**91.71**	**96.59**	**94.15**	98.30	88.97	**86.49**	**91.25**	**91.54**	84.73*	**92.33**	**89.53**	**82.33**	**74.22**	**86.30**

- Visual backbone: XVLM (12M), Swin Transformer [1]
- COGT-XVLM is trained on COCO
- COGT-XVLM+ is trained on CC3M + COCO + Visual Genome

[1] Yan Zeng et al. Multi-grained vision language pre-training: Aligning texts with visual concepts. ICML 2022.

COGT-InstructBLIP

Model	ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
	Relation	**Attribute**	**Avg**	**Add**	**Replace**	**Swap**	**Avg**	**Attribute**	**Object**	**Relation**	**Avg**	**ITT**	**Avg**	
BLIP	59.00	88.00	73.50	.	.	.	.	75.20	82.20	70.50	75.70	.	.	.
BLIP2	41.20	71.30	56.25	.	.	.	.	77.80	84.90	70.60	77.80	.	.	.
InstructBLIP (FlanT5XL)	69.20	50.83	60.02	65.43	72.59	63.41	67.14	56.37	80.33	53.34	63.35	40.33*	26.80*	51.53*
MiniGPT-4	46.90	55.70	51.30	.	.	.	.	71.30	84.20	.	.	.	.	.
GPT-4V	.	.	.	91.68	**93.37**	86.61	90.55	.	.	.	.	.	.	.
LLaVA-1.5-13B	.	.	.	.	.	80.95	.	.	.	.	.	.	.	.
LLaVA-1.5-13B+CRG	.	.	.	.	.	87.90	.	.	.	.	.	.	.	.
LLaVA-1.6-34B	.	.	.	.	.	81.25	.	.	.	.	.	.	.	.
LLaVA-1.6-34B+CRG	.	.	.	.	.	90.75	.	.	.	.	.	.	.	.
BLIP-VisualGPTScore ($\alpha = 0$) †	89.10*	95.30	92.20*	91.00	93.30	**91.00**	91.77*	78.70*	**92.60**	90.80	87.37	.	.	.
BLIP2-VisualGPTScore ($\alpha = 0$) †	90.70	94.30*	92.50	92.70	93.00*	91.24	92.31	73.90	90.10	89.90*	84.63	.	.	.
Cap	86.60	88.90	87.75	98.94	88.21	84.00	90.38	.	.	.	.	.	.	
CapPa	86.70	85.70	86.20	**99.13**	87.67	83.11	89.97	.	.	.	.	.	.	.
COGT-InstructBLIP	87.63	88.93	88.28	98.55*	90.61	88.12	**92.42**	85.77	79.96	89.14	84.96*	72.66	51.26	77.87
COGT-InstructBLIP+	**91.12**	**95.64**	**93.38**	98.45	90.27	88.22*	92.31	**90.80**	85.17*	**92.80**	**89.60**	**83.33**	**70.72**	**85.87**

- Visual backbone: InstructBLIP-flan-t5-xl, ViT g/14 + q-former [1]
- COGT-InstructBLIP is trained on COCO
- COGT-InstructBLIP+ trained on CC3M + COCO + Visual Genome

[1] Wenliang Dai et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. NeurIPS, 2023.

FG-OVD

- FG-OVD [1] was originally proposed to evaluate the ability of open-vocabulary object detectors to discern fine-grained object properties
- In FG-OVD Negative captions are created starting from the object-specific captions by replacing attributes referring to the object's color, material, texture, etc.

✓ *A laptop computer with a grey metal back, featuring a white logo made of metal.*

✗ A laptop computer with a red metal back, featuring a white logo made of metal.
✗ A laptop computer with a grey metal back, featuring a white logo made of glass.
✗ A laptop computer with a dark orange metal back, featuring a white logo made of metal.
✗ A laptop computer with a grey metal back, featuring a white logo made of plastic.
✗ A laptop computer with a grey metal back, featuring a white logo made of crochet
✗ A laptop computer with a grey crochet back, featuring a white logo made of metal.
✗ A laptop computer with a pink metal back, featuring a white logo made of metal.
✗ A laptop computer with a grey leather back, featuring a white logo made of metal.
✗ A laptop computer with a grey metal back, featuring a dark purple logo made or metal.
✗ A laptop computer with a grey metal back, featuring a white logo made of stone.

DAC-LLM ✗

COGT ✓

✓ *A light blue and light grey plastic clock with a text pattern and a black metal hand.*

✗ A light blue and light grey plastic clock with a text pattern and a black fabric hand.
✗ A light blue and light grey fabric clock with a text pattern and a black metal hand.
✗ A light blue and light red plastic clock with a text pattern and a black metal hand.
✗ A light blue and light grey plastic clock with a studded pattern and a black metal hand.
✗ A light blue and light grey plastic clock with a text pattern and a black ceramic hand.
✗ A light blue and light grey plastic clock with a text pattern and a white metal hand.
✗ A light blue and light grey plastic clock with a text pattern and a black wool hand.
✗ A light blue and light grey plastic clock with a text pattern and a yellow metal hand.
✗ A light blue and light grey plastic clock with a striped pattern and a black metal hand.
✗ A light blue and light grey plastic clock with a text pattern and a black crochet hand.

DAC-LLM ✗

COGT ✓

[1] Lorenzo Bianchi et al. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. CVPR 2024

Conclusions

COGT introduces a compositional method using semi-parallel training.

- It leverages an off-the-shelf dependency parser to establish causal relations between words.
- These relations are encoded in a Causal Graphical Model (CGM), which reduces spurious associations in the joint probability distribution.
- This structure enhances data efficiency, making better use of training data and reducing overfitting.
- Experimental results demonstrate that COGT significantly outperforms previous compositional approaches, even those trained on larger datasets

Code is available at https://github.com/aimagelab/COGT

Thank you for your attention

Fiorenzo Parascandolo Nicholas Moratelli Enver Sangineto Lorenzo Baraldi Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy