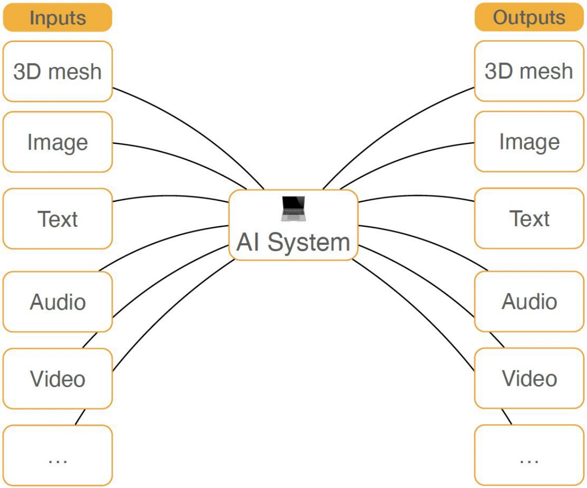
Criteo – Trustworthy AI Symposium

Explainability framework for large multimodal models

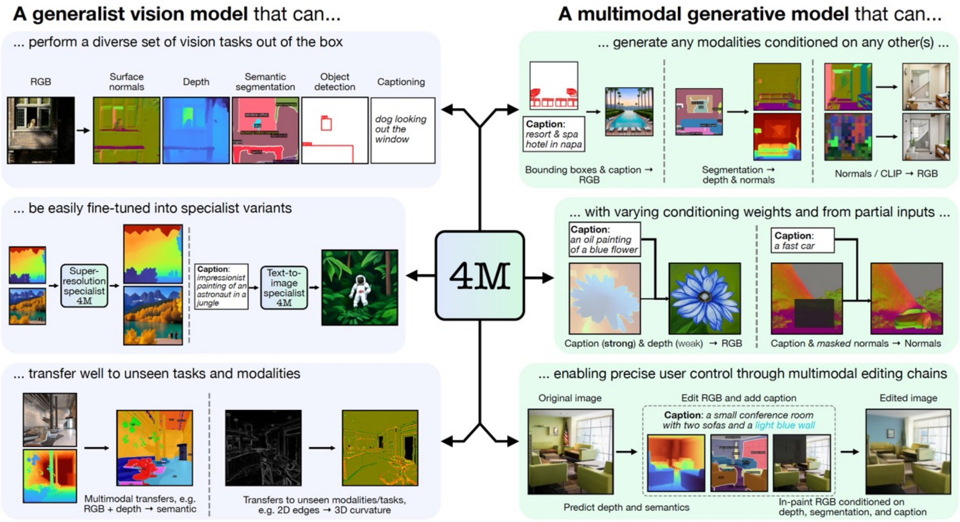
Matthieu Cord
Sorbonne Université, valeo.ai

Collaborators: Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson

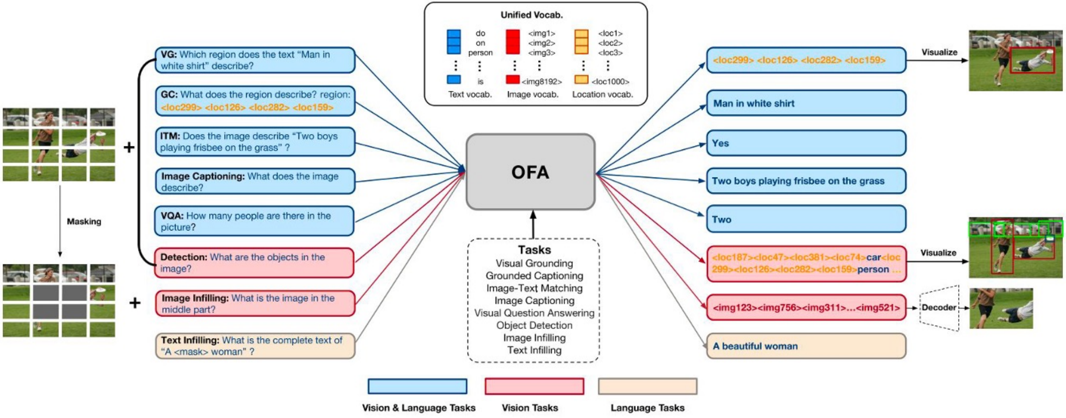
Multimodal learning tasks and models



Multimodal learning tasks and models

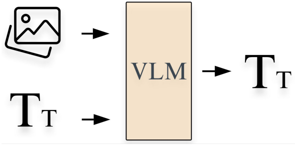


Multimodal learning tasks and models



Multimodal learning tasks and models

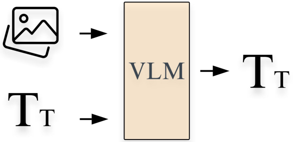
- Which Multimodal models?
  - Vision Encoder + LLM Decoder
  - Image (+ text) as input, textual caption as output

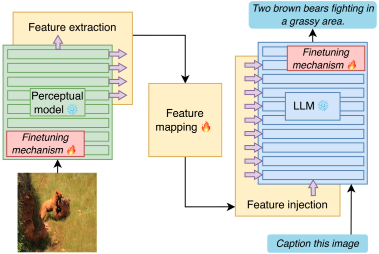


Multimodal learning tasks and models

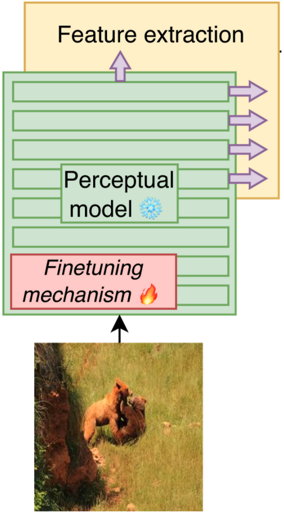
- Which Multimodal models?
  - Vision Encoder + LLM Decoder
  - Image (+ text) as input, textual caption as output

- Focus on Large Multimodal Models (LMMs) processing visual and language data
  $\Longrightarrow$ Popular for solving Visual captioning, question-answering, reasoning tasks

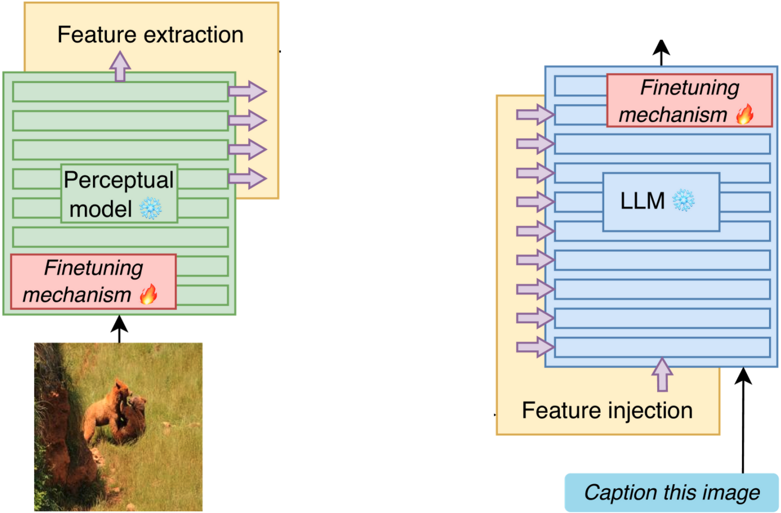




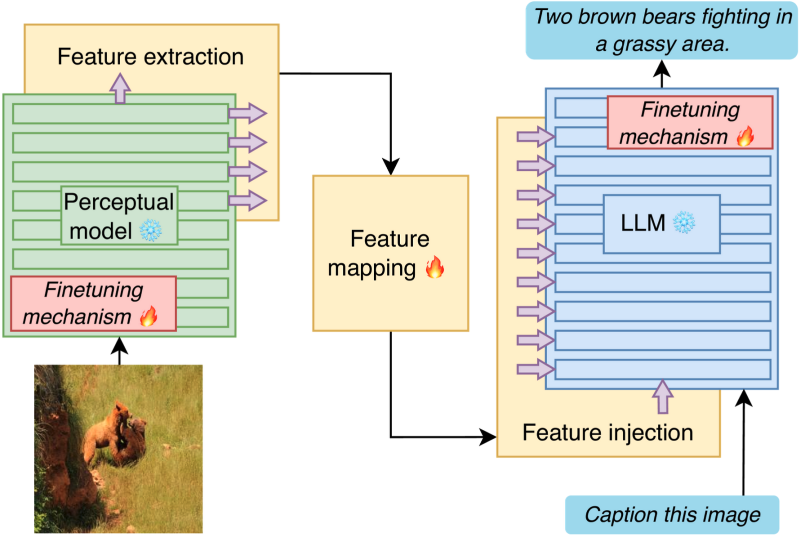
Which Multimodal models?



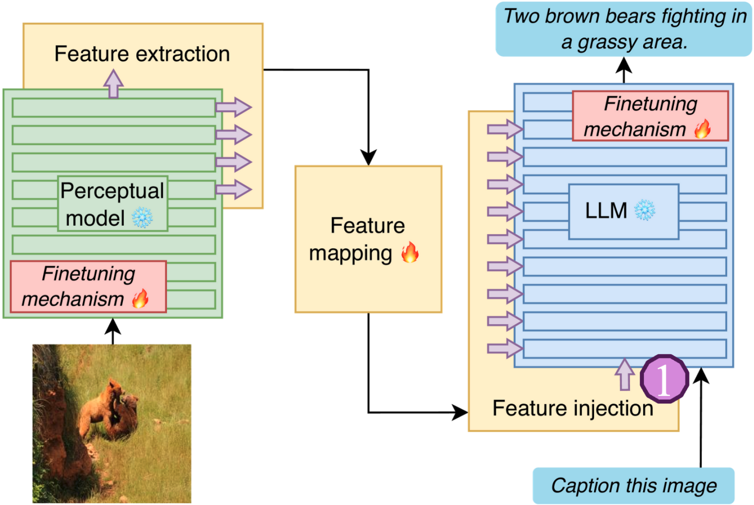
Which Multimodal models?



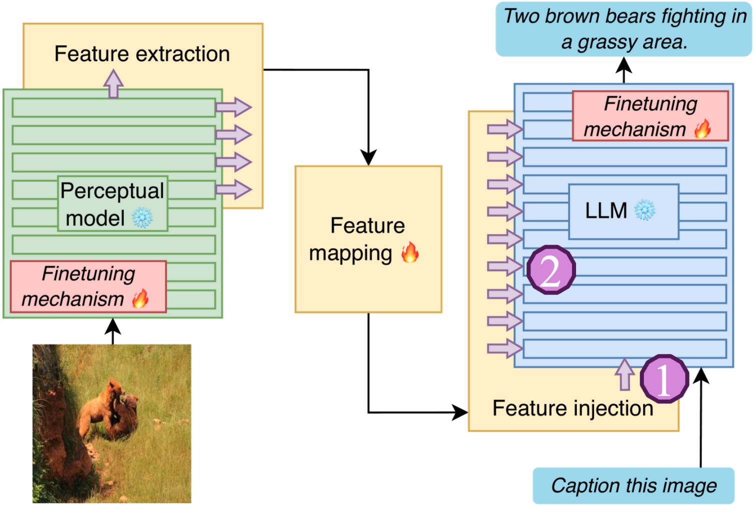
Which Multimodal models?



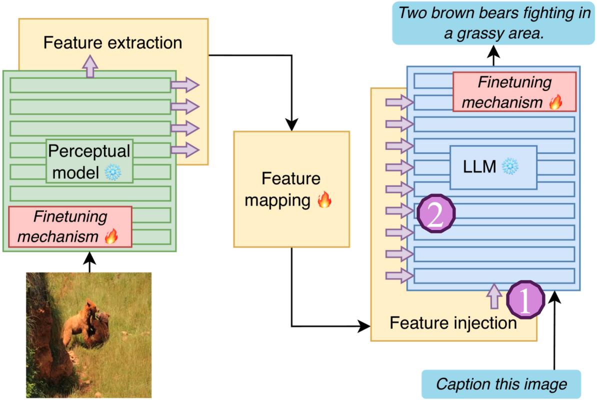
Which Multimodal models?

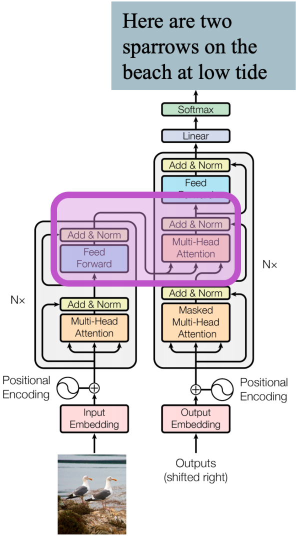


Which Multimodal models?

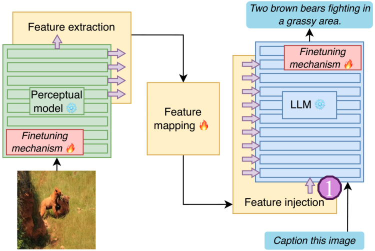


Which Multimodal models?





Which Multimodal models?

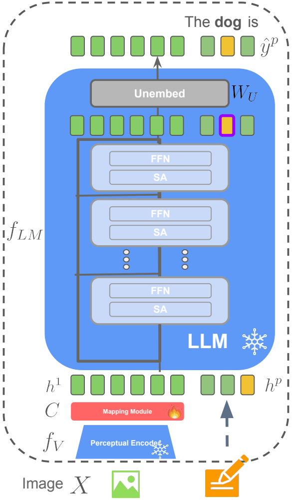


2021	2022	2022	2023	2023	2024
CLIP	**CoCa / FLAVA**	**Flamingo**	**Idefics1 / OBELICS**	**LLaVA**	**Idefics2**
Foundation model for image-text embeddings	Vision-language models with single image conditioning	Interleaved text-images with multimodal web documents	Open-Source reproduction of Flamingo and its web document dataset	Fully autoregressive architecture with frozen pre-trained components	Fully autoregressive architecture, LoRA on pre-trained component, and pooling of visual tokens

CoX-LMM (NeurIPS24): Explaining/Monitoring LMMs

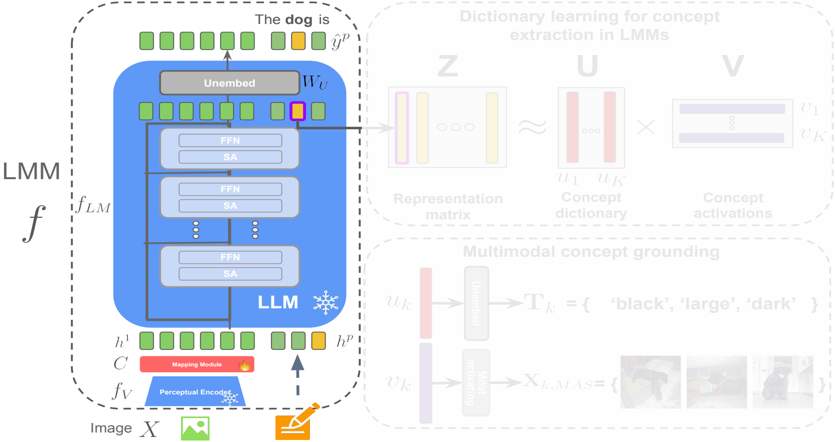
Monitoring LMMs: Supervising, Observing, Tracking, Watching, Overseeing, Surveying, ...

- Pretrained LMM f = Visual encoder (f_V) + Connector (C) + Language model (f_{LM})
- Captioning dataset $\mathcal{S} = \{(X_i, y_i)\}_{i=1}^{N}$. Images $X_i \in \mathcal{X}$ and captions $y_i \subset \mathcal{Y}$
- A token of interest $t \in \mathcal{Y}$ (Eg. 'Dog', 'Cat' etc.)
- **Analysis**: Understand internal representations of f about t in terms of high-level concepts



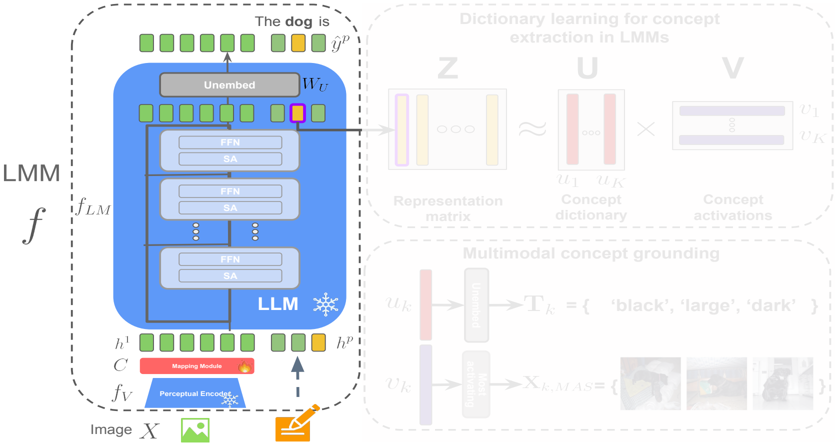
CoX-LMM: A Concept based eXplainability framework for LMMs

Monitoring LMM: CoX-LMM



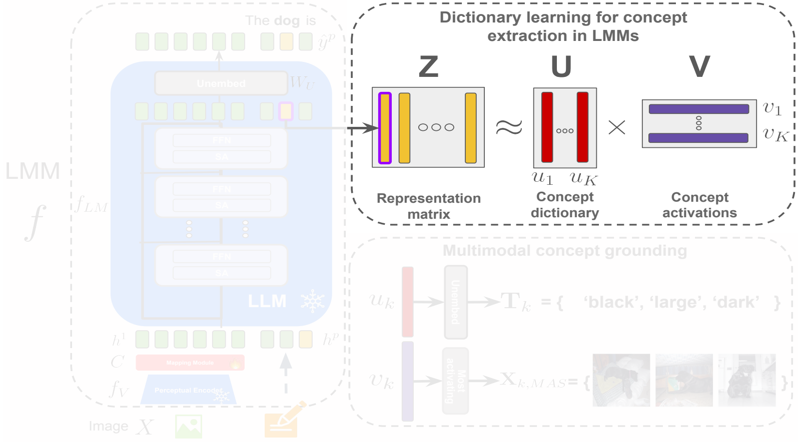
- Input to f_{LM} - Concatenated sequence of tokens: (1) Visual tokens $C(f_V(X))$, (2) textual tokens previously predicted by f_{LM}
- Caption predicted by f_{LM} trained for next-token prediction task

Monitoring LMM: CoX-LMM



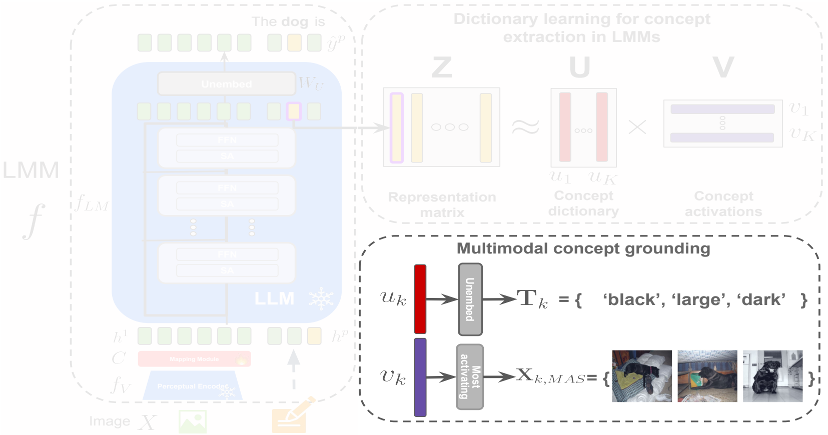
- Extract residual stream representations of t from f for a relevant set of M images $\mathbf{X}$
- Collect all such B-dimensional representations as columns of matrix $\mathbf{Z} \in \mathbb{R}^{B \times M}$

Monitoring LMM: CoX-LMM



- Dictionary learning for concept extraction. Semi-NMF optimization:
  $\mathbf{U}^*, \mathbf{V}^* = \arg\min_{\mathbf{U},\mathbf{V}} \; ||\mathbf{Z} - \mathbf{U}\mathbf{V}||_F^2 + \lambda ||\mathbf{V}||_1 \quad s.t. \; \mathbf{V} \geq 0,$ and $||u_k||_2 \leq 1 \; \forall k \in \{1, ..., K\}$
- Columns of $\mathbf{U}^* \in \mathbb{R}^{B \times K}$ – concept vectors. Rows of $\mathbf{V}^* \in \mathbb{R}^{K \times M}$ – concept activations

CoX-LMM: Multimodal concept grounding!

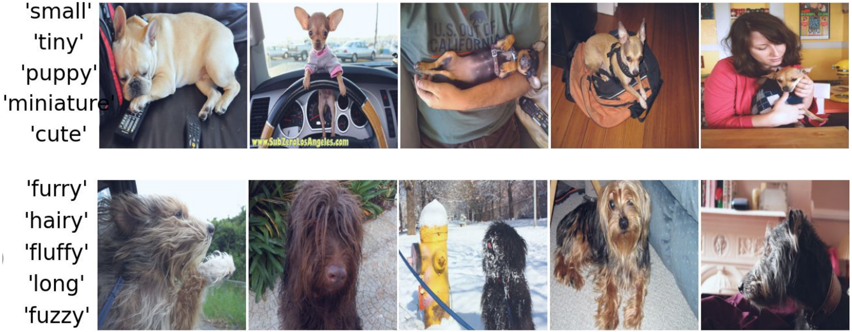


- **Text grounding**: Decode concept vector u_k with f_{LM} head and extract top tokens
- **Visual grounding**: Extract most activating samples for u_k (via activations v_k)

Example multimodal concepts

Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!

- Visual: Most activating images of u_k from $\mathbf{X}$ (via $v_k \in \mathbb{R}^M$) $\rightarrow \mathbf{X}_{k,MAS}$
- Textual: unembedding matrix W_U decode u_k and extract the most probable tokens $\rightarrow \mathbf{T}_k$



Example multimodal concepts

Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!

Example multimodal concepts

Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!

Example multimodal concepts

Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!

'cat'
'kitten'
'tiger'
'rabbit'
'dog'

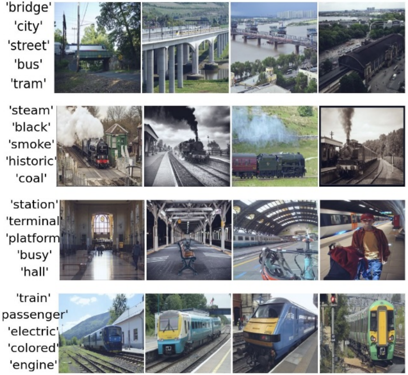
'herd'
'sheep'
'flock'
'farm'
'shepherd'

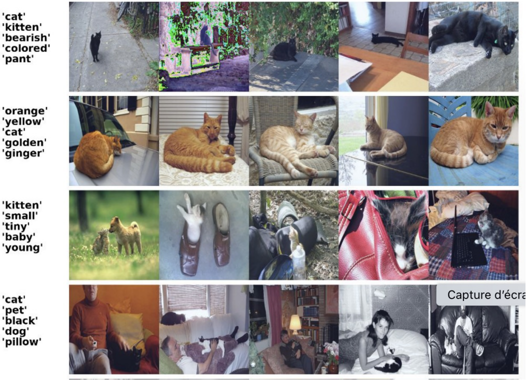
Example multimodal concepts

Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!

Example multimodal concepts

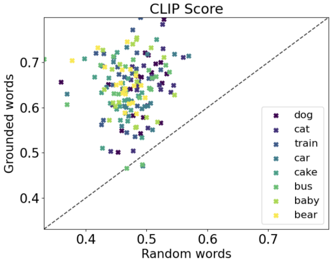
Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!

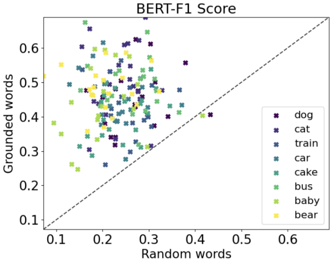


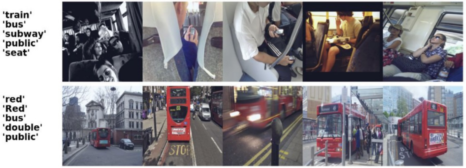


Multimodal grounding evaluation

- CLIPScore or BERTScore (for captions) between $\mathbf{X}_{k,MAS}$ and $\mathbf{T}_k$ (vs $\mathbf{R}_k$).
- Averaged over all MAS samples or all their associated captions.

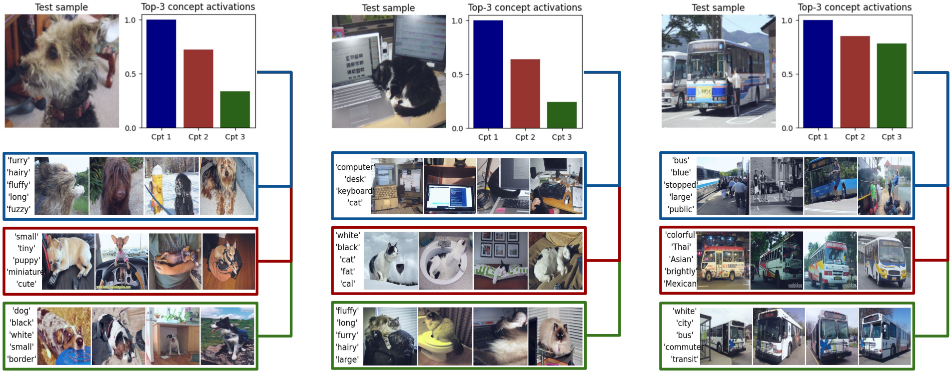






Using the concept dictionary

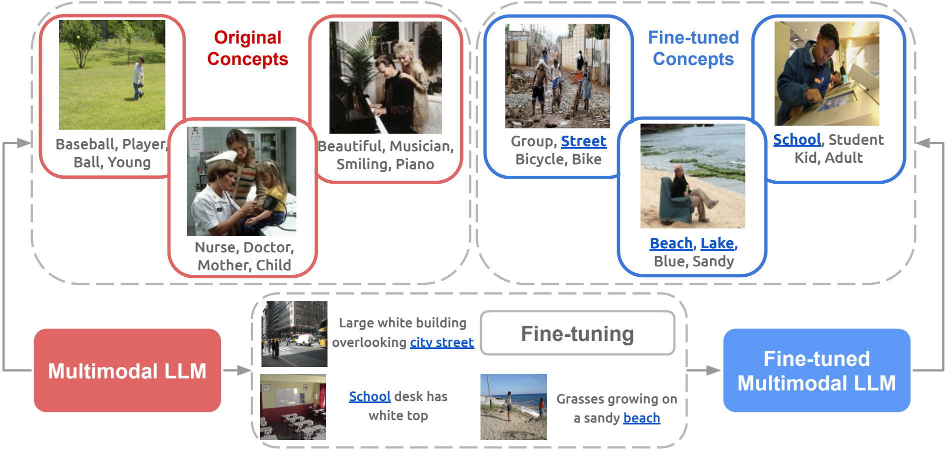
- For a new image X where $t \in f(X)$, extract z_X and compute the projection on $\mathbf{U}^*$, $v(X) = \arg\min_{v \geq 0} ||z_X - \mathbf{U}^* v||_2^2 + \lambda ||v||_1$
- **Most activating concepts**: From $v(X)$ we can extract the concept activations with largest magnitudes, $\tilde{u}(X)$



Using the concept dictionary

What happens if we fine-tune the LMM?

- How do concepts encoded with the initial model change when we fine-tune it?
- Is it possible to manipulate the output of an LMM without fine-tuning it?

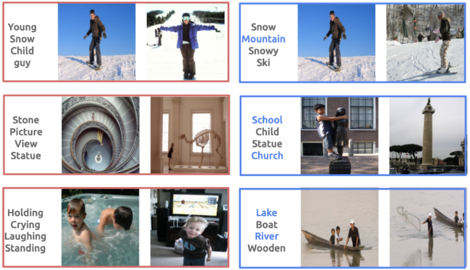


Change of concepts

- matching function $m : i \to j^*$, for $u_i^a \in U^a$:

$$m(i) = \underset{u_j^b \in U_b}{\operatorname{argmax}} \cos(u_i^a, u_j^b)$$

- Concepts are refined, emerged, or diminished.



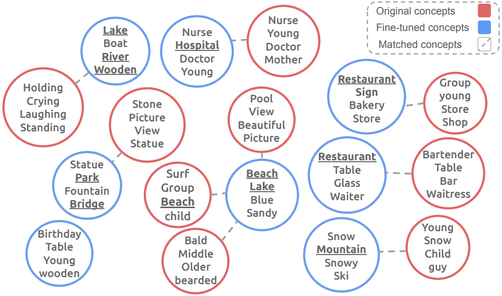


Figure: **Concepts text grounding change after fine-tuning.** Illustration of text grounding for concepts (TOI = person) from f^a and their match from f^b, after fine-tuning to focus more on places.

Concept recovery

1. Associate each concept u_k^a in the original model with a subset of samples:

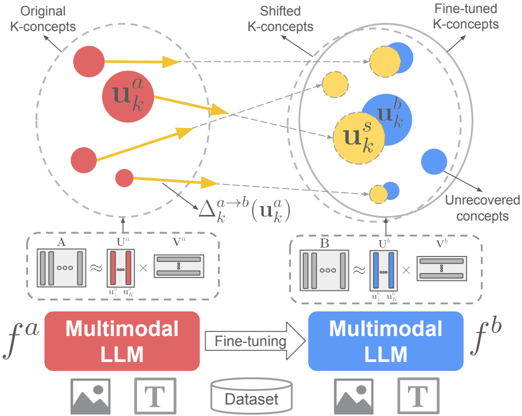
$$A_k = \{m \mid k = \underset{i}{\operatorname{argmax}} \, |v_i^a(x_m)|\}.$$

2. For each sample, $x_m, m \in A_k$ define $\delta_m^{a\to b} = b_m - a_m$ as an individual shift vector, and then aggregate them for one concept:

$$\Delta_k^{a\to b}(u_k^a) = \frac{1}{|A_k|} \sum_{m \in A_k} \delta_m^{a\to b}$$

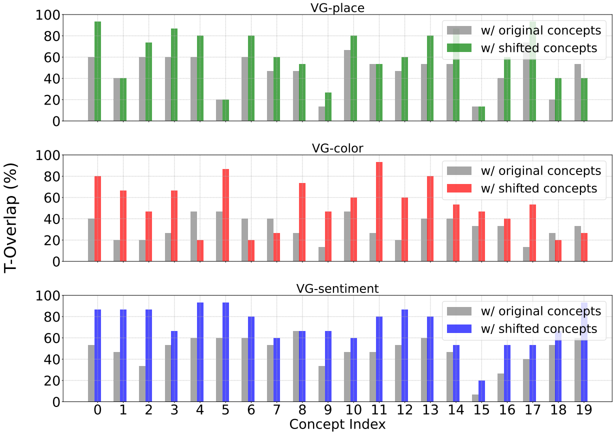
3. Shift an original concept with the shift vector:

$$u_k^s = u_k^a + \alpha \, \Delta_k^{a\to b}(u_k^a)$$



Concept recovery

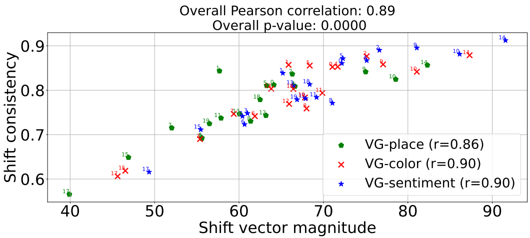
- Recovery metric : T-Overlap$(u, u') = 100 \times \frac{\left|T_{\text{words}}(u) \cap T_{\text{words}}(u')\right|}{\left|T_{\text{words}}(u)\right|}$
- Comparison between T-Overlap$(u_k^a, u_{m(k)}^b)$ and T-Overlap$(u_k^s = u_k^a + \alpha\, \Delta_k^{a \to b}(u_k^a), u_{m(k)}^b)$



Shift consistency, steering

- Shift consistency: how aligned are individual shift vectors corresponding to a concept

$$\text{Consistency}(u_k^a) = \frac{1}{|A_k|}\sum_{m \in A_k} \cos(\delta_m, \Delta_k^{a \to b}(u_k^a))$$

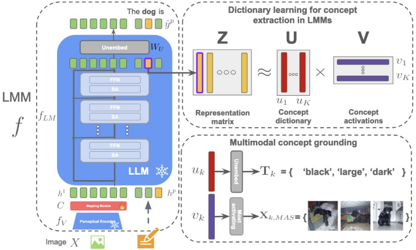


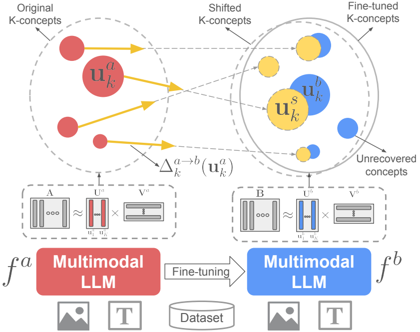
- Steering the captioning:

Conclusion

- Monitoring LMMs with multimodal concepts
- Analyzing MLLMs' internal representations after fine-tuning:
  - Demonstrated that post-fine-tuning concepts can often be recovered from the original model
  - Steering model behavior by modifying features directly, without additional training

→ Can steering vectors define/learn a more general steering function? The revanche of REFT on PEFT!





Thank you for your attention!

Matthieu Cord
Sorbonne Université, valeo.ai

Collaborators: Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson

Project webpage: https://jayneelparekh.github.io/LMM_Concept_Explainability/

Code: https://github.com/mshukor/xl-vlms