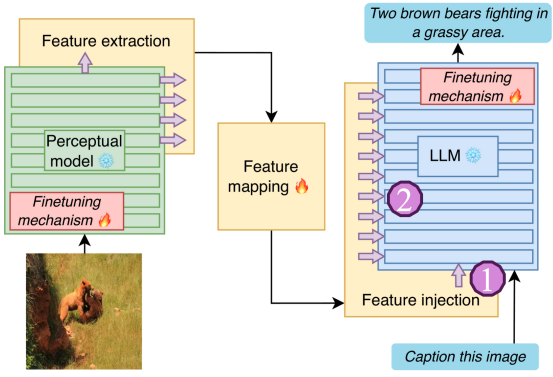
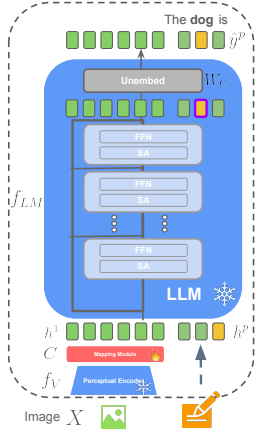
Explaining/Monitoring LMMs





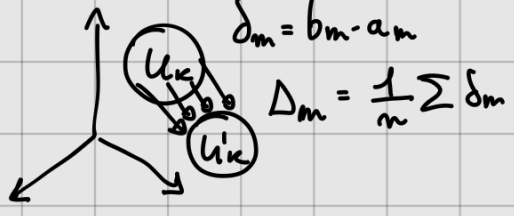
Monitoring LLMs

- LMM $f = f_{LM}(C(f_V))$
  - f_{LM}: language model, C: connector, f_V: visual encoder
- Understand f internal representation of a given token "dog" using CoX-LLM
  - one example → B samples in latent space → K clusters u_K
    - characterized both in terms of text and of images

Fine-tuned LLMs

After fine-tuning concepts change (refined / emerged / diminished)

- stronger for new concepts, inefficient for some previously learnt concept (forgotten)
- compute concepts shift



Steering

Linear modification to align the output in latent space

- refusal, truthful, ...

P2S → L2S

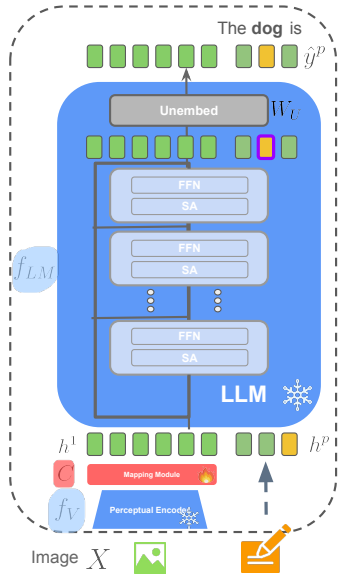
P2S: fixed steering direction

L2S: steering direction input dependent

Explaining/Monitoring LMMs

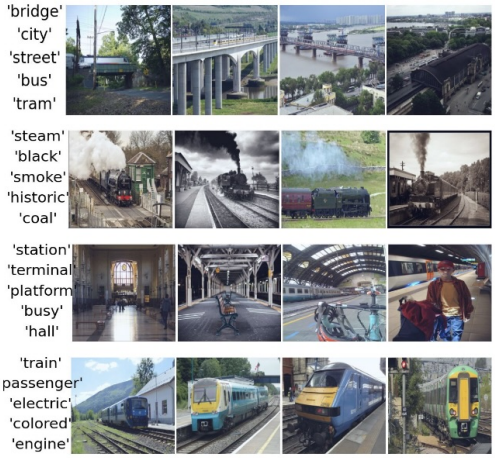
- Pretrained LMM f = Visual encoder (f_V) + Connector (C) + Language model (f_{LM})
- Captioning dataset $\mathcal{S} = \{(X_i, y_i)\}_{i=1}^{N}$. Images $X_i \in \mathcal{X}$ and captions $y_i \subset \mathcal{Y}$
- A token of interest $t \in \mathcal{Y}$ (Eg. 'Dog', 'Cat' etc.)
- **Analysis**: Understand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

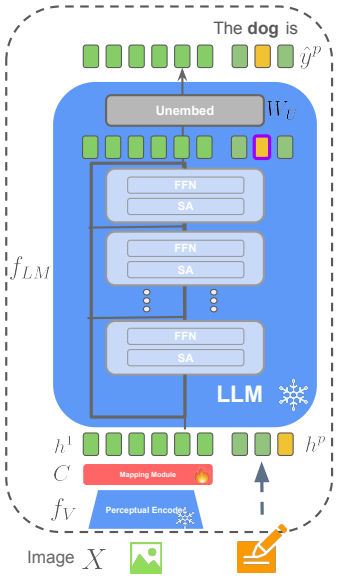
Concept based eXplainability framework for LMMs



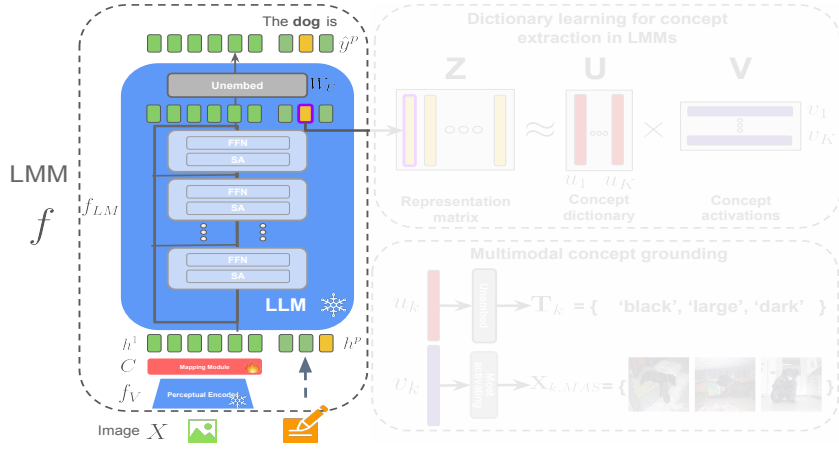
Explaining/Monitoring LMMs

For token of interest t 'Train', can we provide a multimodal concept analysis? such as:



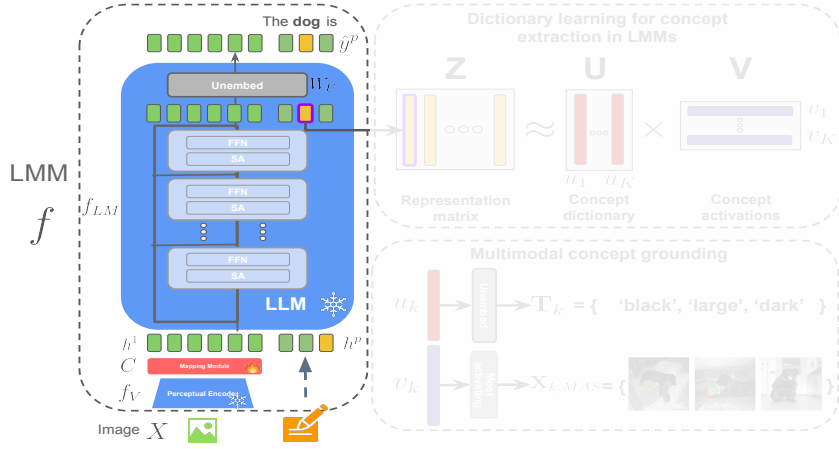


Monitoring LMM



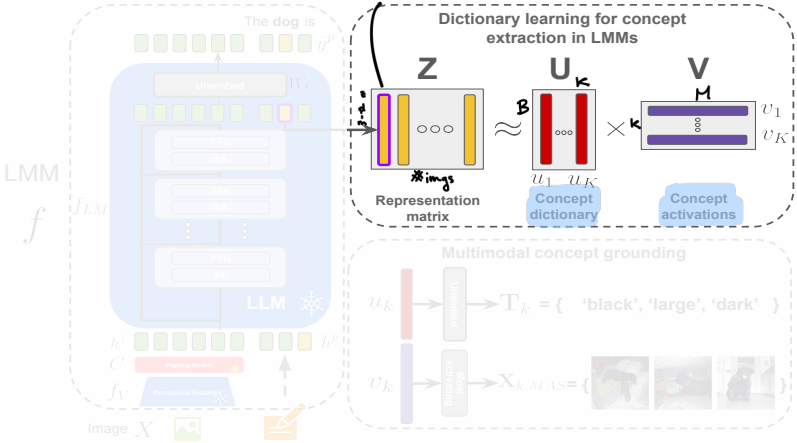
- Input to f_{LM} - Concatenated sequence of tokens: (1) Visual tokens $C(f_V(X))$, (2) textual tokens previously predicted by f_{LM}
- Caption predicted by f_{LM} trained for next-token prediction task

Monitoring LMM



- 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



- Dictionary learning for concept extraction. Semi-NMF optimization:
  $\mathbf{U}^*, \mathbf{V}^* = \arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{Z}-\mathbf{UV}||_F^2+\lambda||\mathbf{V}||_1 \quad s.t.\ \mathbf{V} \geq 0,$ 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

a L1 penalty (LASSO) → drive most v_k to 0, use the fewest v_k. Explanation are sparse

CoX-LMM: Multimodal concept grounding!

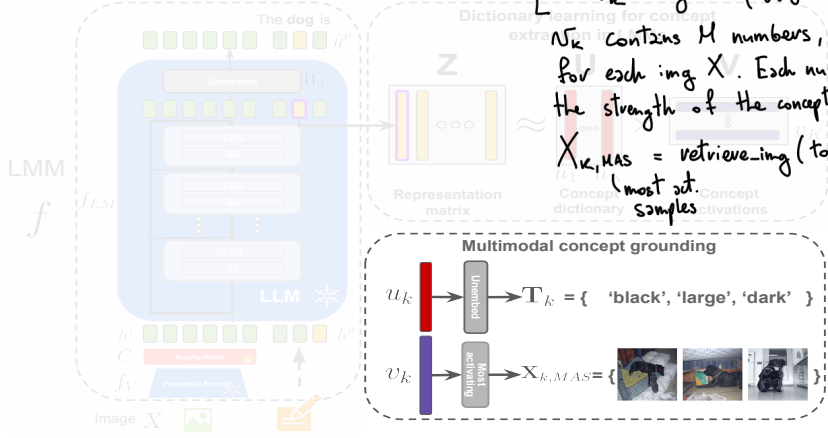
W_U unembedding matrix

$T_k = \text{argmax}(W_U \cdot v_k)$

v_k contains M numbers, one for each img X. Each num is the strength of the concept activation

$X_{k,MAS} = \text{retrieve\_img}(\text{top}_N(v_k))$

↳ most act. samples



- **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)

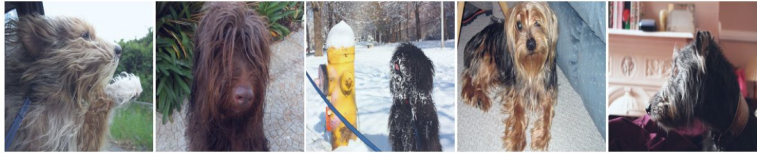
Example multimodal concepts

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 $\rightarrow \mathbf{T}_k$

'small'
'tiny'
'puppy'
'miniature'
'cute'

'furry'
'hairy'
'fluffy'
'long'
'fuzzy'



Example multimodal concepts

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

'black'
'large'
'dark'
'big'
'close'

'brown'
'large'
'dog'
'tan'
'golden'

Example multimodal concepts

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

top_5_tokens$(W_v \cdot v_k)$

retrieve_img$(\text{top}_5(r_k))$

'dog'
'running'
'black'
'play'
'grass'

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'

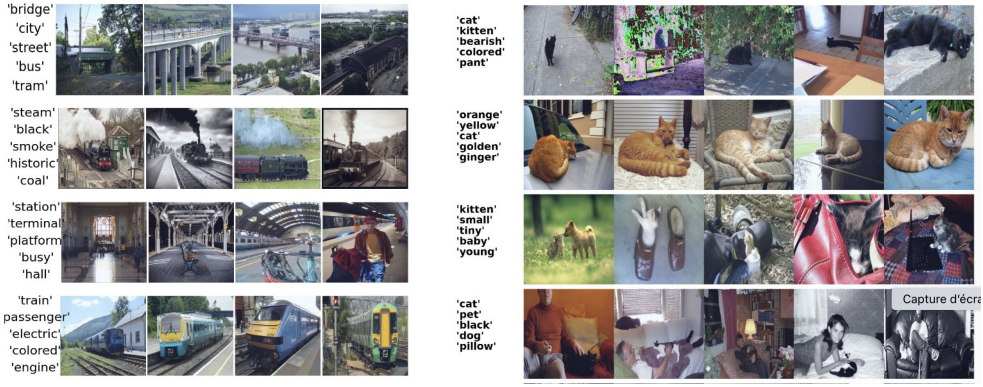
Example multimodal concepts

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

'dog'
'sausage'
'hot'
'sandwich'
'plate'

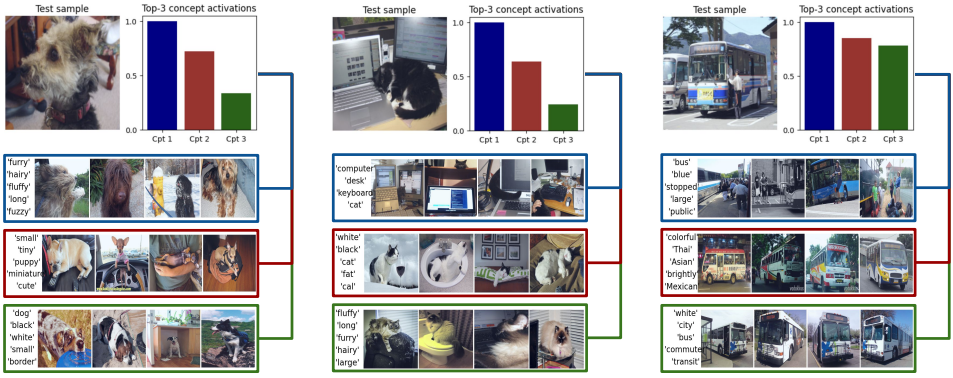
Example multimodal concepts

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



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? It destroys previous U,V building new ones

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

Yes, instead of changing W, just manipulate v for a certain concept u

u_k = "dog"

v_k = [0, 0.3, 0.1, 0] here we have a dog

$\tilde{v}_k$ = [0.9, 0, 0, 0] then reproject back

$z_{steered} = U \cdot \tilde{v}_k$