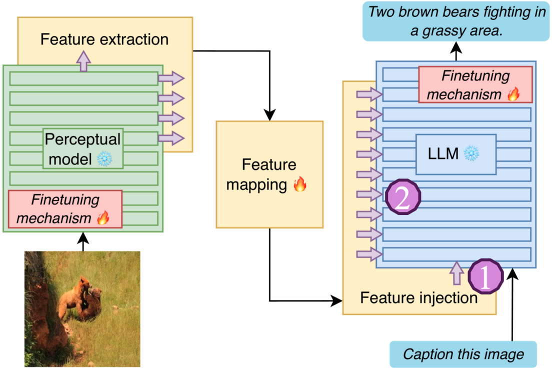
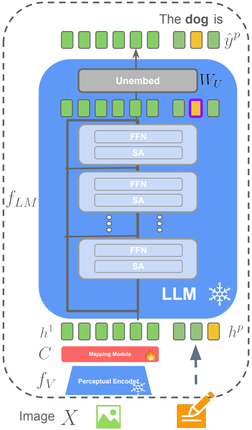
Explaining/Monitoring LMMs

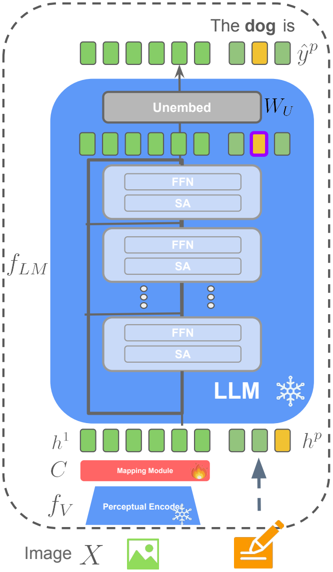




Explaining/Monitoring LMMs

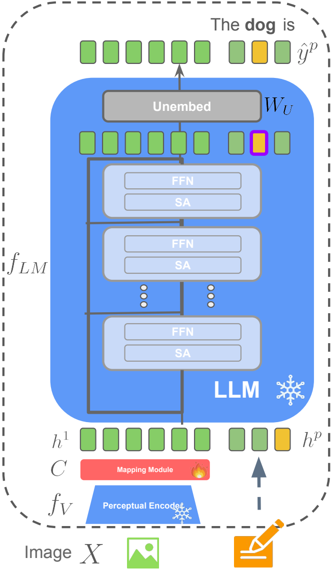
- Pretrained LMM f = Visual encoder (f_V) + Connector (C) + Language model (f_{LM})
- Captioning dataset $\mathcal{S} = \{(X_i, y_i)\}_{i=1}^N$. Images $X_i \in \mathcal{X}$ and captions $y_i \subset \mathcal{Y}$
- A token of interest $t \in \mathcal{Y}$ (Eg. 'Dog', 'Cat' etc.)
- **Analysis**: Understand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

Concept based eXplainability framework for LMMs

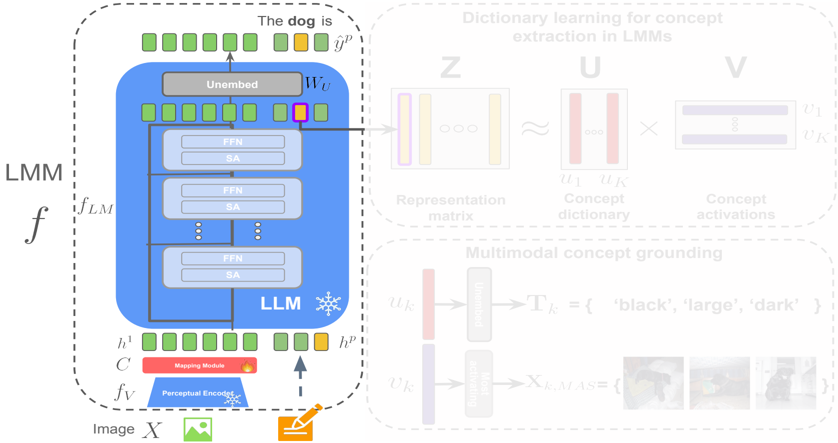


Explaining/Monitoring LMMs

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

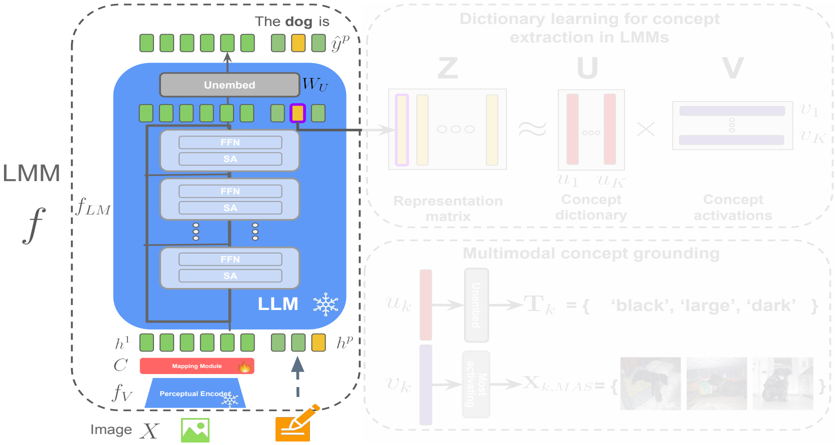


Monitoring LMM



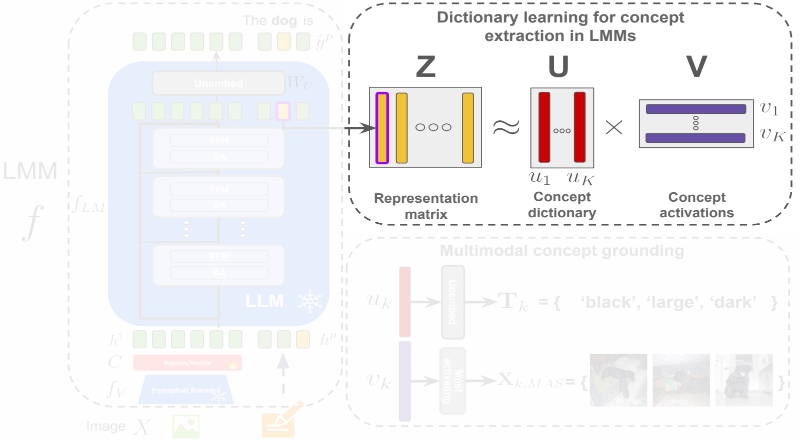
- Input to f_{LM} - Concatenated sequence of tokens: (1) Visual tokens $C(f_V(X))$, (2) textual tokens previously predicted by f_{LM}
- Caption predicted by f_{LM} trained for next-token prediction task

Monitoring LMM



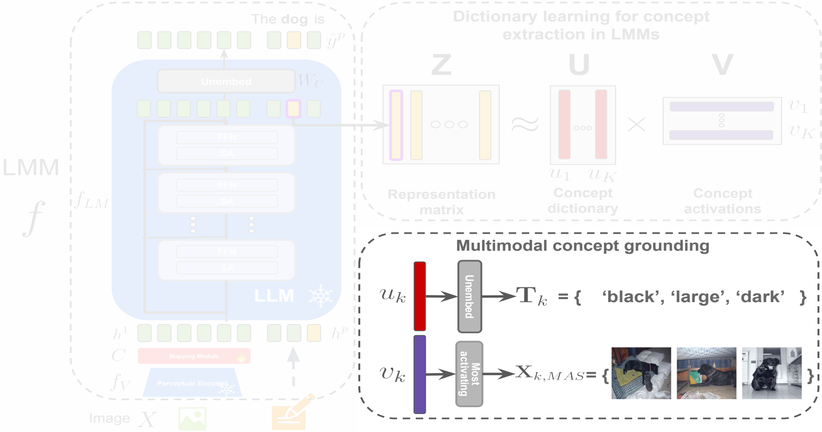
- Extract residual stream representations of t from f for a relevant set of M images $\mathbf{X}$
- Collect all such B-dimensional representations as columns of matrix $\mathbf{Z} \in \mathbb{R}^{B \times M}$

Monitoring LMM



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

'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!

Example multimodal concepts

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

Example multimodal concepts

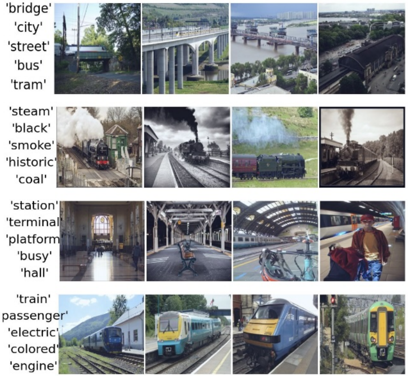
Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!

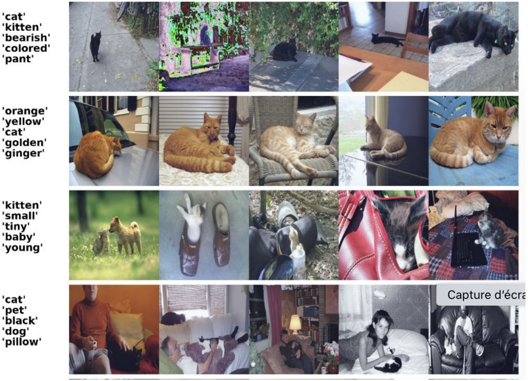
Example multimodal concepts

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

Example multimodal concepts

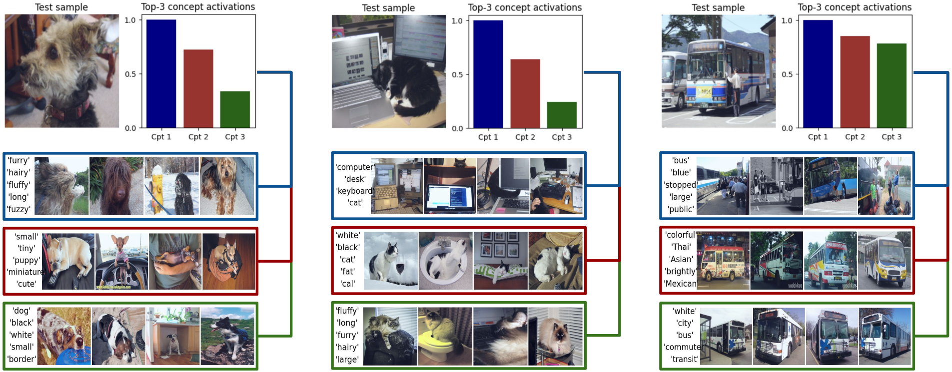
Multimodal concepts: $u_k \in \mathbf{U}^*$ simultaneously grounded in both vision and text!





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?

- 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?