Energy-Based Concept Bottleneck Models: Unifying Prediction, Concept Intervention, and Probabilistic Interpretations

Xinyue Xu, Yi Qin, Lu Mi, Hao Wang*, Xiaomeng Li*

18 Apr 2024

* Equal advising

Concept-Based Models

[Koh 2020]

Input image x

Concepts c

Output label y

DNN

Black head: 0.9

Yellow breast: 0.2

Small: 0.8

Black and White Warbler

After Intervention

DNN

Black head: 0.9

Yellow breast: 1.0

Small: 0.8

Intervention

Domain expert

Kentucky Warble

Current Limitations

- **Interpretability**

  Cannot effectively quantify the intricate relationships between various concepts and class labels.

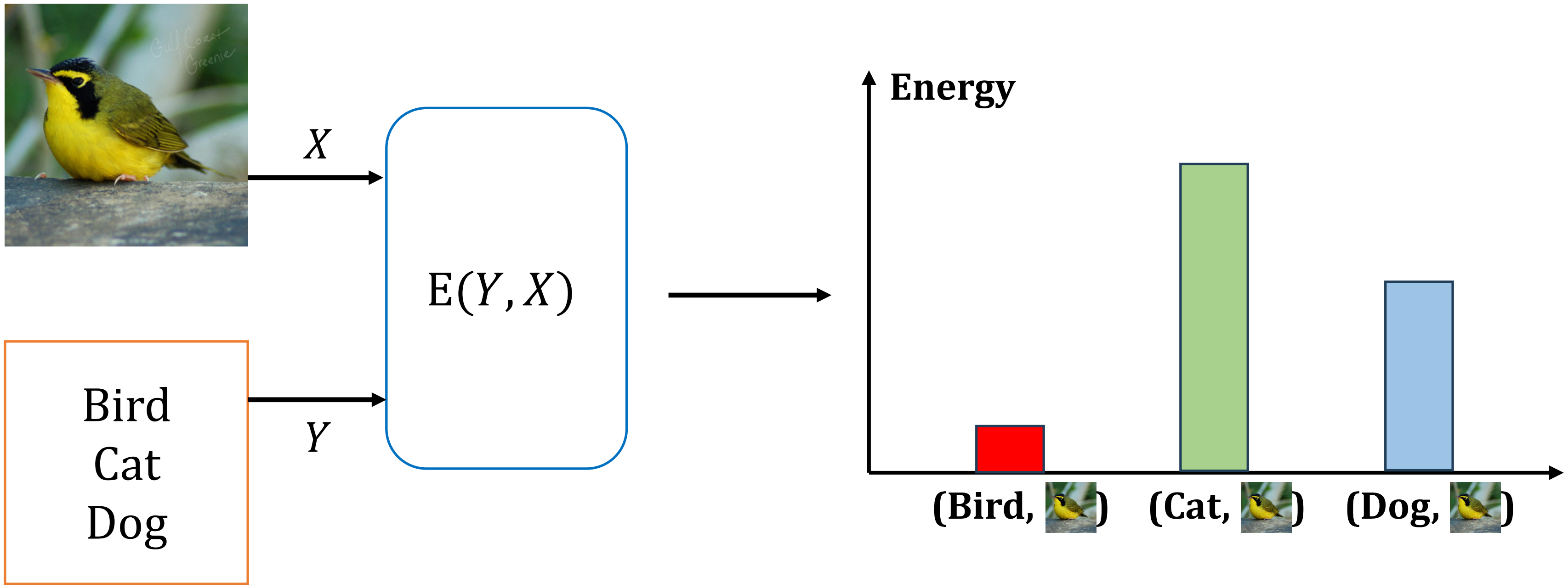
- **Intervention**

  Struggle to account for the complex interactions among concepts.

- **Performance**

  Suffer from a trade-off between model performance and interpretability.

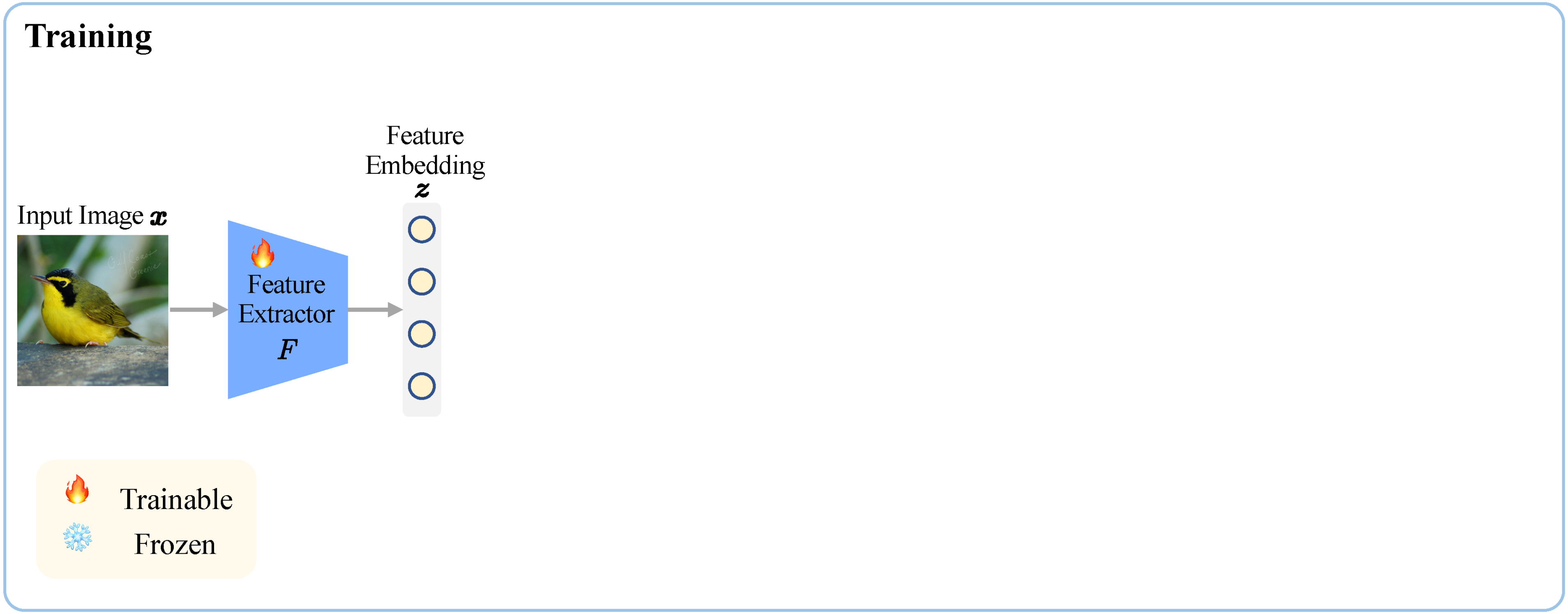
Energy-Based Models



For energy networks, lower energy $\boldsymbol{E}$ indicates better compatibility (e.g., $\boldsymbol{E}(\boldsymbol{x}, y_{gt}) = 0$).

Our Method (ECBM): Feature Extractor

Given the input x and a candidate label y, the feature extractor F first compute the features $z = F(x)$.

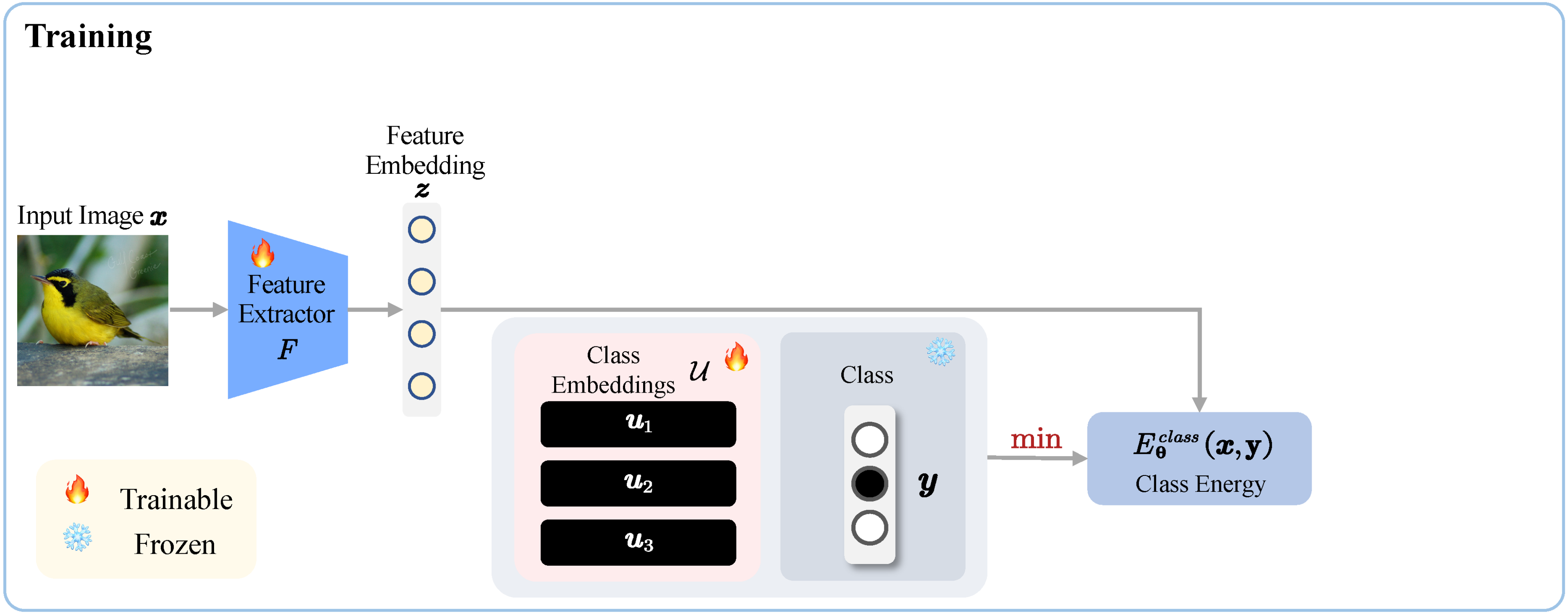


ECBM: Class Energy Network $E_{\theta}^{class}(x, y)$

Measure the compatibility between input x and class label y.

$$E_{\theta}^{class}(x, y) = G_{zu}(z, u)$$

$$\mathcal{L}_{class}(\boldsymbol{x}, \boldsymbol{y}) = E_{\boldsymbol{\theta}}^{class}(\boldsymbol{x}, \boldsymbol{y}) + \log\left(\sum_{m=1}^{M} e^{-E_{\boldsymbol{\theta}}^{class}(\boldsymbol{x}, \boldsymbol{y}_m)}\right). \quad (1)$$

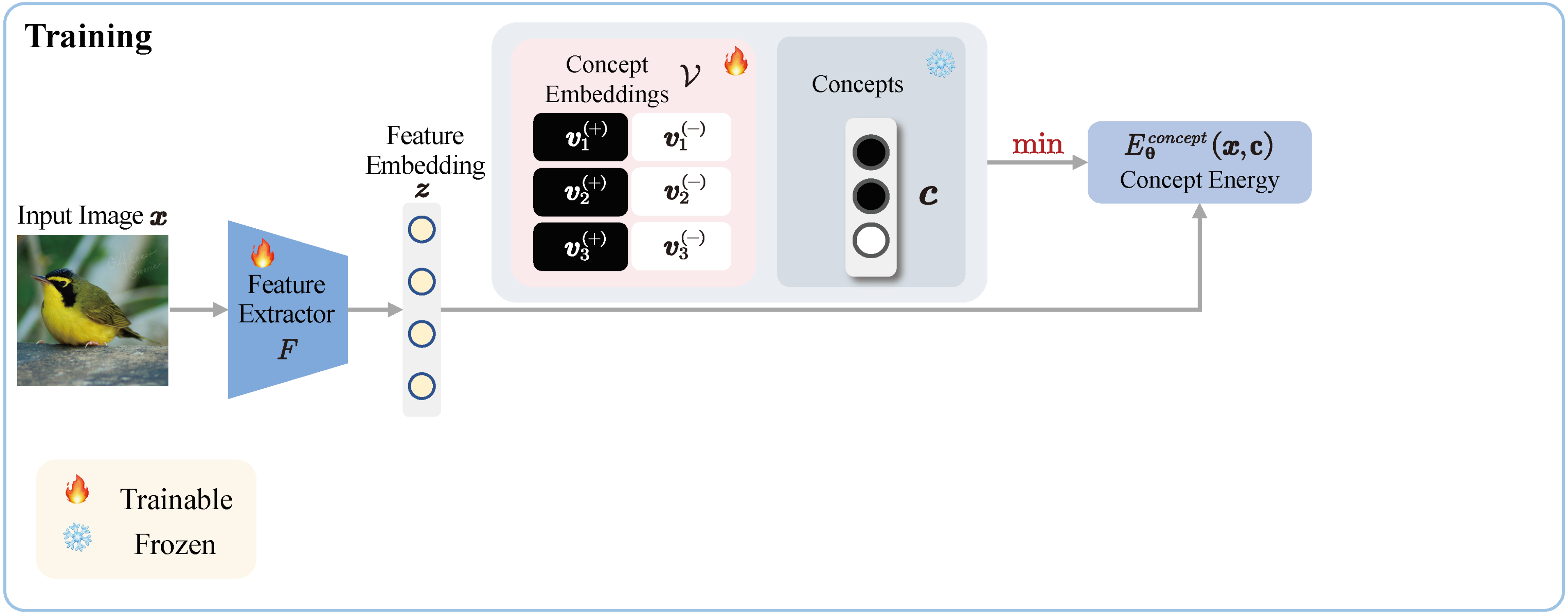


ECBM: Concept Energy Network $E_{\theta}^{concept}(x, c)$

Measure the compatibility between input x and the K concepts c.

$$E_{\theta}^{concept}(x, c_k) = G_{zv}(z, v_k)$$

$$\mathcal{L}_{concept}^{(k)}(\boldsymbol{x}, c_k) = E_{\boldsymbol{\theta}}^{concept}(\boldsymbol{x}, c_k) + \log\Big(\sum\nolimits_{c_k \in \{0,1\}} e^{-E_{\boldsymbol{\theta}}^{concept}(\boldsymbol{x}, c_k)}\Big). \quad (2)$$

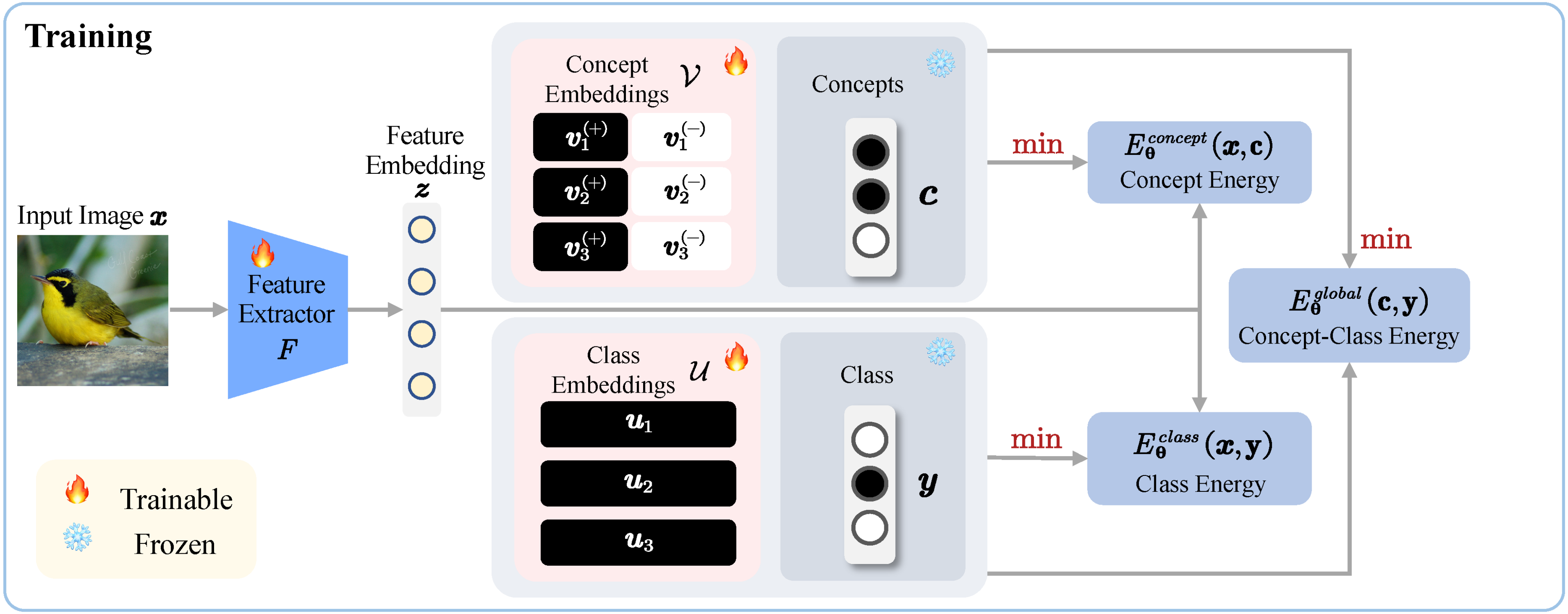


ECBM: Global Energy Network $E_{\theta}^{global}(c, y)$

Measure the compatibility between the K concepts c and class label y.

$$E_{\theta}^{global}(c, y) = G_{vu}\left([v_k]_{k=1}^{K}, u\right)$$

$$\mathcal{L}_{global}(\boldsymbol{c}, \boldsymbol{y}) = E_{\boldsymbol{\theta}}^{global}(\boldsymbol{c}, \boldsymbol{y}) + \log\left(\sum\nolimits_{m=1, \boldsymbol{c}' \in \mathcal{C}}^{M} e^{-E_{\boldsymbol{\theta}}^{global}(\boldsymbol{c}', \boldsymbol{y}_m)}\right). \quad (3)$$

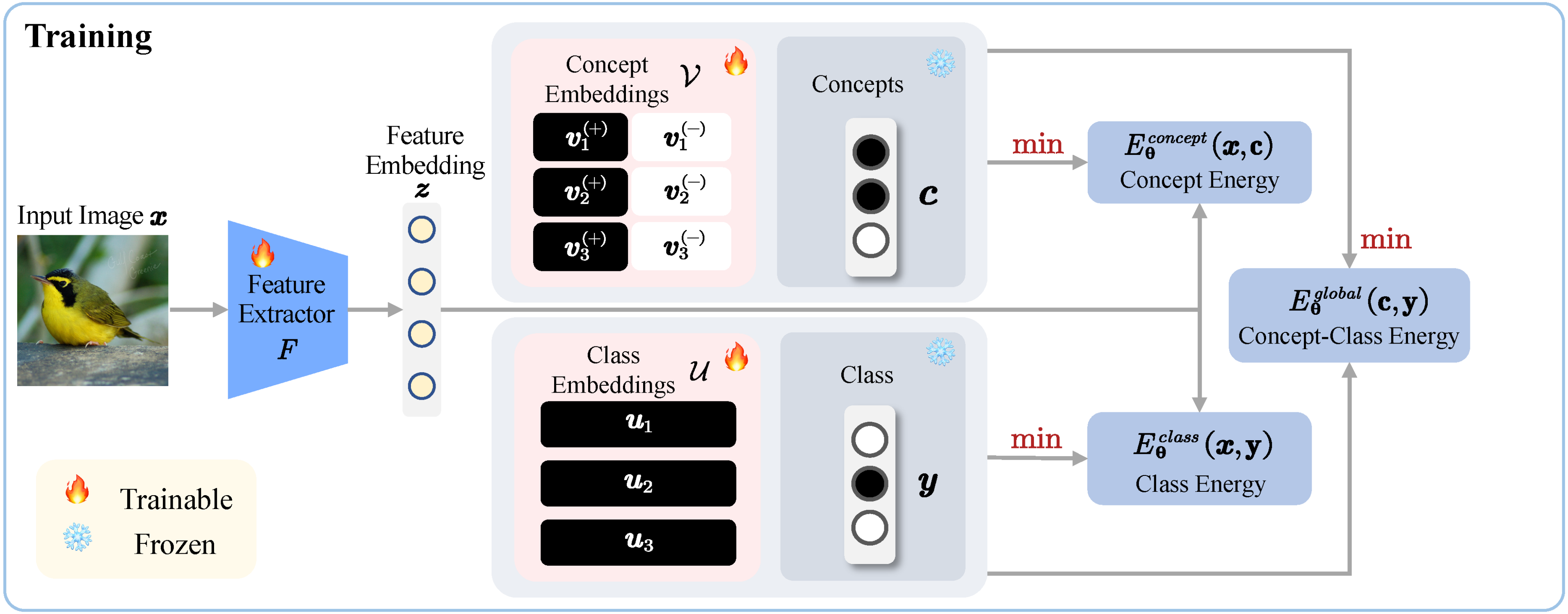


Training Phase: Minimize Loss

ECBM is trained by minimizing the following total loss function:

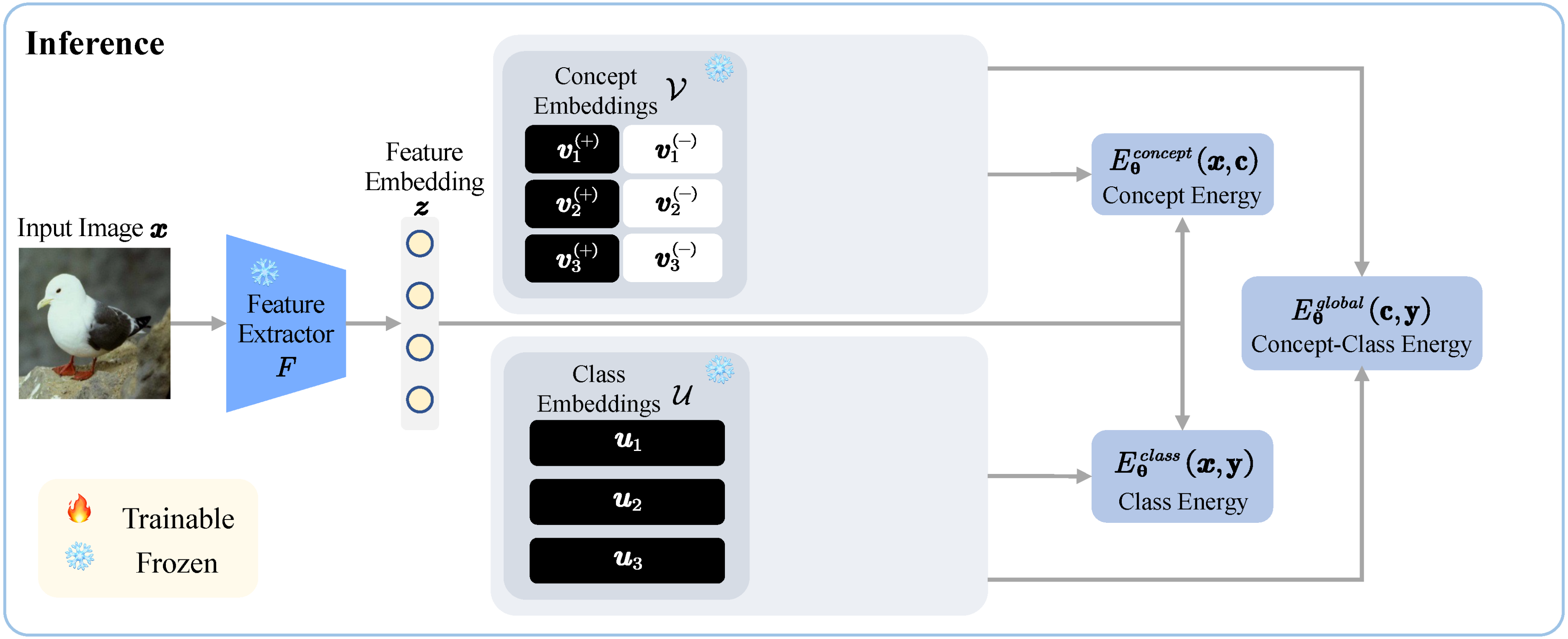
$$\mathcal{L}_{total}(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}) = \mathcal{L}_{class}(\boldsymbol{x}, \boldsymbol{y}) + \lambda_c \mathcal{L}_{concept}(\boldsymbol{x}, \boldsymbol{c}) + \lambda_g \mathcal{L}_{global}(\boldsymbol{c}, \boldsymbol{y}), \tag{4}$$

$$\mathcal{L}_{total}^{all} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{c},\boldsymbol{y}) \sim p_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{c},\boldsymbol{y})}[\mathcal{L}_{total}(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y})]. \tag{5}$$



Inference Phase : Freeze Parameters

To predict $\boldsymbol{c}$ and $\boldsymbol{y}$ given the input $\boldsymbol{x}$, we **freeze** the feature extractor F and the energy network parameters $\boldsymbol{\theta}$.

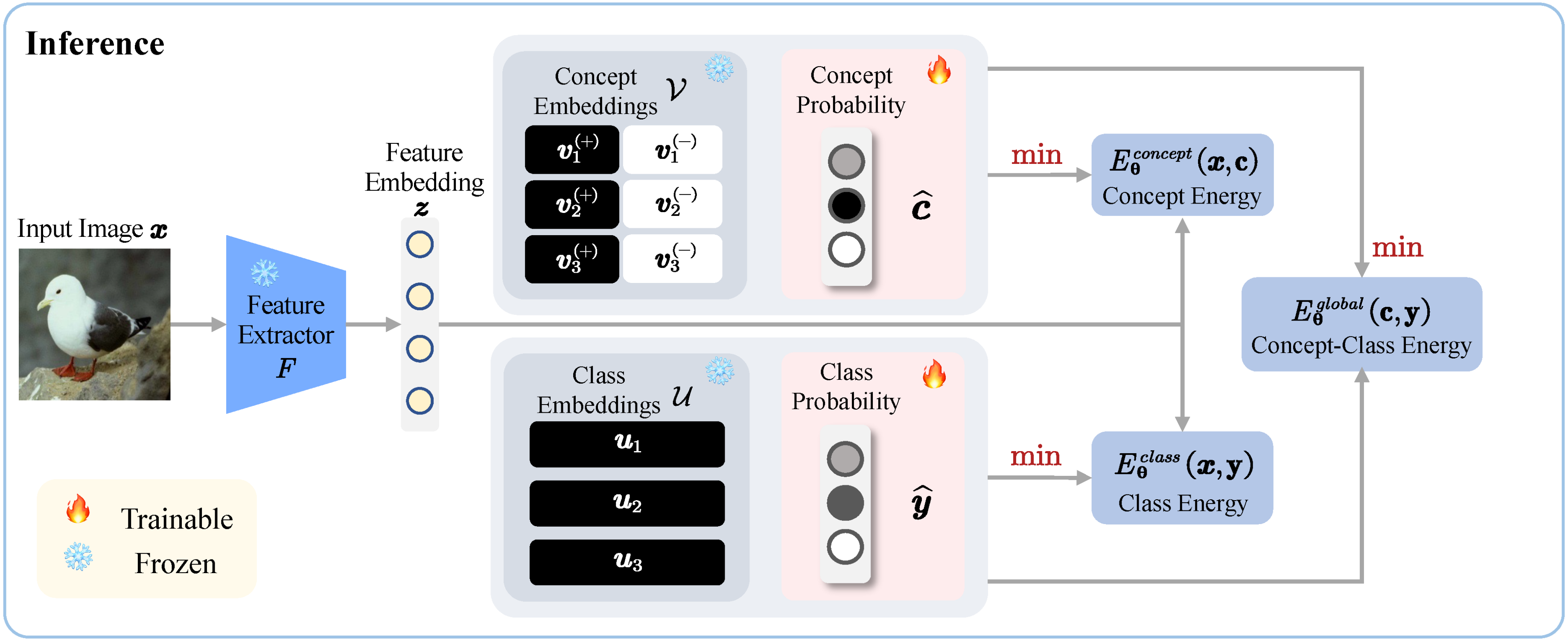


Inference Phase: Search Optimum

Search for the optimal prediction of concepts $\widehat{\boldsymbol{c}}$ and the class label $\widehat{\boldsymbol{y}}$ as follows:

$$\arg\min_{\widehat{\boldsymbol{c}},\widehat{\boldsymbol{y}}} \; \mathcal{L}_{class}(\boldsymbol{x},\widehat{\boldsymbol{y}}) + \lambda_c \mathcal{L}_{concept}(\boldsymbol{x},\widehat{\boldsymbol{c}}) + \lambda_g \mathcal{L}_{global}(\widehat{\boldsymbol{c}},\widehat{\boldsymbol{y}}), \tag{6}$$

$$E_{\boldsymbol{\theta}}^{joint}(\boldsymbol{x},\boldsymbol{c},\boldsymbol{y}) \triangleq E_{\boldsymbol{\theta}}^{class}(\boldsymbol{x},\boldsymbol{y}) + \lambda_c E_{\boldsymbol{\theta}}^{concept}(\boldsymbol{x},\boldsymbol{c}) + \lambda_g E_{\boldsymbol{\theta}}^{global}(\boldsymbol{c},\boldsymbol{y}). \tag{7}$$



Experimental Results

Model \ Data	CUB			CelebA			AWA2		
Metric	Concept	Overall Concept	Class	Concept	Overall Concept	Class	Concept	Overall Concept	Class
CBM	0.964	0.364	0.759	0.837	0.381	0.246	**0.979**	0.803	0.907
ProbCBM*	0.946	0.360	0.718	0.867	0.473	0.299	0.959	0.719	0.880
PCBM	-	-	0.635	-	-		-	-	
CEM	0.965	0.396	0.796	0.867	0.457	0.330	0.978	0.796	0.908
ECBM	**0.973**	**0.713**	**0.812**	**0.876**	**0.478**	**0.343**	**0.979**	**0.854**	**0.912**

- **Slightly** outperform others in terms of ***concept accuracy***.

Experimental Results

Model \ Data	CUB			CelebA			AWA2		
Metric	Concept	Overall Concept	Class	Concept	Overall Concept	Class	Concept	Overall Concept	Class
CBM	0.964	0.364	0.759	0.837	0.381	0.246	**0.979**	0.803	0.907
ProbCBM*	0.946	0.360	0.718	0.867	0.473	0.299	0.959	0.719	0.880
PCBM	-	-	0.635	-	-		-	-	
CEM	0.965	0.396	0.796	0.867	0.457	0.330	0.978	0.796	0.908
ECBM	**0.973**	**0.713**	**0.812**	**0.876**	**0.478**	**0.343**	**0.979**	**0.854**	**0.912**

- **Slightly** outperform others in terms of ***concept accuracy***.
- Successfully capture the ***interaction (and correlation)*** among the concepts.

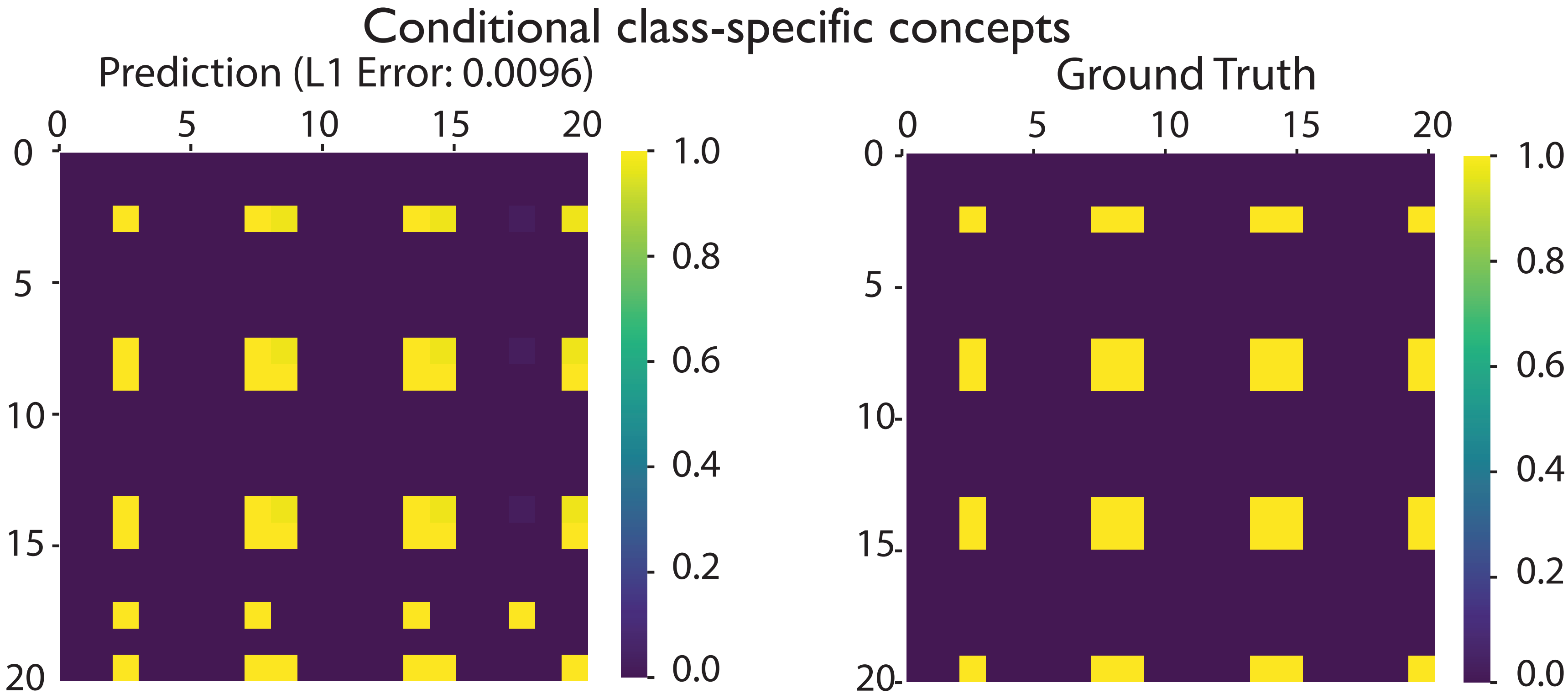
  Significantly outperforms other methods in terms of ***overall concept accuracy***.

Experimental Results

Model \ Data	CUB			CelebA			AWA2		
Metric	Concept	Overall Concept	Class	Concept	Overall Concept	Class	Concept	Overall Concept	Class
CBM	0.964	0.364	0.759	0.837	0.381	0.246	**0.979**	0.803	0.907
ProbCBM*	0.946	0.360	0.718	0.867	0.473	0.299	0.959	0.719	0.880
PCBM	-	-	0.635	-	-		-	-	
CEM	0.965	0.396	0.796	0.867	0.457	0.330	0.978	0.796	0.908
ECBM	**0.973**	**0.713**	**0.812**	**0.876**	**0.478**	**0.343**	**0.979**	**0.854**	**0.912**

- **Slightly** outperform others in terms of ***concept accuracy***.
- Successfully capture the ***interaction (and correlation)*** among the concepts.
    Significantly outperforms other methods in terms of ***overall concept accuracy***.
- Outperform the state-of-the-art on class accuracy.

Conditional Interpretation



$$\mathrm{p}(c_k = 1 | c_{k'} = 1, \boldsymbol{y} = \text{Black and White Warble})$$

Conclusion

- Propose the **first general method - ECBM,** to unify:
  - Concept correction
  - Conditional interpretation
  - Concept-based prediction
- **Under a unified energy formulation,** compute arbitrary conditional probabilities.
- **Significantly** outperform the state-of-the-art on **real-world** datasets.

Paper

Code