EQ-CBM: A Probabilistic Concept Bottleneck with Energy-based Models and Quantized Vectors

Sangwon Kim[1][0000−0002−7452−3897], Dasom Ahn[2][0009−0009−4123−072X],
Byoung Chul Ko[2][0000−0002−7284−0768], In-su Jang[1][0000−0002−0468−4193], and
Kwang-Ju Kim[1,*][0000−0001−8458−4506]

[1] ETRI, South Korea
{eddiekim,jef1015,kwangju}@etri.re.kr
[2] Keimyung University, South Korea
tommydasomahn@gmail.com, niceko@kmu.ac.kr
* Corresponding author

Abstract. The demand for reliable AI systems has intensified the need for interpretable deep neural networks. Concept bottleneck models (CBMs) have gained attention as an effective approach by leveraging human-understandable concepts to enhance interpretability. However, existing CBMs face challenges due to deterministic concept encoding and reliance on inconsistent concepts, leading to inaccuracies. We propose EQ-CBM, a novel framework that enhances CBMs through probabilistic concept encoding using energy-based models (EBMs) with quantized concept activation vectors (qCAVs). EQ-CBM effectively captures uncertainties, thereby improving prediction reliability and accuracy. By employing qCAVs, our method selects homogeneous vectors during concept encoding, enabling more decisive task performance and facilitating higher levels of human intervention. Empirical results using benchmark datasets demonstrate that our approach outperforms the state-of-the-art in both concept and task accuracy.

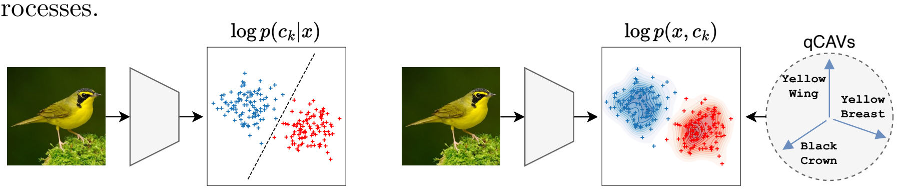
Keywords: Concept Bottleneck Model · Energy-based Model · Vector Quantization

1 Introduction

With the increasing demand for reliable AI, explaining deep neural networks (DNNs) has gained significant attention across various research fields. For decades, post-hoc explanation methods [33,2,23,30,35] have been the mainstay due to their clarity and compatibility. However, these methods fall short in fully interpreting black-box DNNs as they provide explanations detached from the intrinsic decision-making processes [24,19] of models. Recently, concept-based model interpretation has emerged as a promising alternative. Kim *et al.* [15] define concepts as discriminative vectors necessary for a model to understand objects, presented as orthogonal vectors known as concept activation vectors (CAVs). Building on this, concept bottleneck models (CBMs) [20,6,40,17,42,3,32,12,34,18] have been proposed to explain the decision-making processes of DNNs using human-understandable concepts. For example, these concepts include `feather color`, `body shape`, or `beak length`, representing distinct attributes of a bird (object) in an image. Various CBMs express these concepts as scores, representing the probability of each concept's presence in the input. CBMs leverage these interpretable concepts for final task prediction without relying on auxiliary features, thereby enhancing interpretability and transparency.

Despite their potential, previous CBMs face limitations in real-world scenarios due to their deterministic concept encoding approach, which uses parametric mappings from input features to concepts. This can lead to inaccuracies, often mapping different concepts to the same latent variables. For instance, CEM [6] shows minimal performance changes in human intervention tests without random intervention learning. Moreover, the decision-making process in CBMs can rely on inconsistent concepts, leading to less discriminative concept utilization and impeding effective human intervention. Incorrect concept encoding from complex inputs further constrains task performance.

To address these challenges, we propose a framework that enhances CBMs through probabilistic concept encoding using **E**nergy-based models (EBMs) with **Q**uantized concept activation vectors (qCAVs), referred to as **EQ-CBM**. EBMs enable robust probabilistic inference by modeling the joint energy-concept landscape for each concept. By incorporating qCAVs, our method selects homogeneous vectors during concept encoding, thereby improving task accuracy and interpretability across various images. This enhancement also facilitates greater human intervention, increasing the model's reliability in complex decision-making processes.



(a) Deterministic Concept Encoding **(b)** Probabilistic Concept Encoding with qCAVs

Fig. 1: Comparison of concept encoding methods: (a) Deterministic concept encoding, and (b) probabilistic concept encoding using qCAVs.

As illustrated in Fig. 1, our approach contrasts with previous deterministic concept encoding. In deterministic encoding, concepts are directly mapped from latent vectors to a fixed representation, often leading to less accurate and interpretable results. In contrast, our approach employs energy-based models to infer the relationship between latent vectors and qCAVs, allowing for a more nuanced and probabilistically robust representation of concepts. This approach not only enhances interpretability but also ensures more consistent and reliable concept encoding across various scenarios. Our contributions include:

- Propose EQ-CBM, a novel framework that enhances CBMs through probabilistic concept encoding using EBMs with qCAVs.
- Introduce robust decision-making using qCAVs to select relevant concepts, improving performance and interpretability.

- Validate EQ-CBM's superiority in interpretability and accuracy through extensive experiments across multiple datasets.

2 Backgrounds

Concept Bottleneck Models [20,6,40,17,42,3,32,12,34,18] were designed to enhance the interpretability of DNNs by leveraging human-understandable concepts. These models transform raw input images into a set of intermediate concept representations, which are then used to make final tasks without relying on additional image features. To explain CBMs in detail, we first define the notations. A dataset $\mathcal{D}$ consists of N triplets, each containing an image x, a ground truth concept label set $\mathbf{C}^*$, and a ground truth class label y^*: $\mathcal{D} = (x, \mathbf{C}^*, y^*)^N$. The ground truth concept label set $\mathbf{C}^*$ includes concept labels for K individual concepts, $\mathbf{C}^* \in \{0,1\}^K$. CBMs typically comprises three main components:

1. **Backbone network** ($f : x \to z$) extracts a latent vector z from the input image x.
2. **Concept encoder** ($g : z \to \mathbf{C}$) maps the latent vector z to a set of concepts $\mathbf{C}$.
3. **Downstream layer** ($h : \mathbf{C} \to y$) predicts the final class y using only the predicted concepts $\mathbf{C}$.

To maintain structural transparency in our model, we employ a lightweight backbone network, ResNet34 [13], for f, and a single fully-connected layer for h. This configuration ensures that the model's decision-making process remains interpretable and transparent. By using these interpretable concepts for final classification tasks, CBMs inherently promote transparency and clarity in model predictions. This transparency allows for human intervention to correct model failures or misalignments between concepts and the final task.

Energy-based Models [26,7,43,4,29,5,11,8,16,41,9] are probabilistic frameworks designed to represent complex distributions. The core principle of EBMs is to associate an energy score with each possible state of variables, where lower energy scores correspond to more probable states. This approach provides a flexible and expressive means to capture the underlying dependencies within the data.

EBMs employ an energy function $E_\theta(x)$, which is parameterized by a compact multi-layer perceptron and maps each variable state to a scalar energy score. This function encapsulates the interactions and dependencies among the variables. More likely states are assigned lower energy scores, while less likely states receive higher scores. The probability of x in EBMs is determined by Boltzmann distribution:

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)}, \quad Z(\theta) = \int_x \exp(-E_\theta(x)), \tag{1}$$

where $Z(\theta)$ is a partition function, which is crucial for normalizing the distribution. It sums the contributions of all possible states, allowing the distribution to be properly normalized. However, calculating $Z(\theta)$ directly is often intractable due to the vast number of possible states.

To address this, Kullback-Leibler (KL) divergence is typically minimized to approximate $p_\theta(x)$ to the true data distribution $p_{\mathcal{D}}(x)$. This involves maximizing the expected log-likelihood of p_θ:

$$\max_\theta \mathbb{E}_{p_{\mathcal{D}}}[\log p_\theta(x)] \tag{2}$$

The gradient of the log-likelihood w.r.t. the parameters θ is given by:

$$\frac{\partial \log p_\theta(x)}{\partial \theta} = \mathbb{E}_{p_\theta(x')}\big[\frac{\partial E_\theta(x')}{\partial \theta}\big] - \frac{\partial E_\theta(x)}{\partial \theta} \tag{3}$$

This gradient can be derived as follows:

$$\begin{aligned}
\log p_\theta(x) &= \log \frac{\exp(-E_\theta(x))}{Z(\theta)} = \log[\exp(-E_\theta(x))] - \log[Z(\theta)] && (4)\\
&= -\log Z(\theta) - E_\theta(x) && (5)\\
\nabla_\theta \log p_\theta(x) &= -\frac{1}{Z(\theta)} \nabla_\theta Z(\theta) - \nabla_\theta E_\theta(x) && (6)\\
&= -\frac{1}{Z(\theta)} \nabla_\theta \int_x \exp(-E_\theta(x)) - \nabla_\theta E_\theta(x) && (7)\\
&= \frac{1}{Z(\theta)} \int_{x'} \exp(-E_\theta(x')) - \nabla_\theta E_\theta(x) && (8)\\
&= \int_x^{\prime} \frac{\exp(-E_\theta(x'))}{Z(\theta)} \nabla_\theta E_\theta(x') - \nabla_\theta E_\theta(x) && (9)\\
&= \mathbb{E}_{p_\theta(x')}[\nabla_\theta E_\theta(x')] - \nabla_\theta E_\theta(x) && (10)
\end{aligned}$$

Direct sampling from $p_\theta(x)$ to obtain negative sample x' is often impractical due to the intractability of the partition function $Z(\theta)$. To address this challenge, sampling methods such as Markov Chain Monte Carlo (MCMC) [26,27,28,10] are utilized. MCMC facilitates approximate inference by sampling from the distribution without the need to explicitly compute $Z(\theta)$.

Previous works have employed methods such as Gibbs sampling [1,14,31] and Stochastic Gradient Langevin Dynamics (SGLD) [25,45] to approximate the true distribution, enabling efficient learning and inference in EBMs. SGLD combines gradient descent with Langevin dynamics to sample from the distribution. The SGLD updates are given by:

$$x_0 \sim p_0(x), \quad x_{t+1} = x_t - \frac{\gamma}{2} \frac{\partial E_\theta(x_t)}{\partial x_t} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \gamma), \tag{11}$$

where γ is the step size and ϵ is Gaussian noise. Learning in EBMs involves adjusting the parameters of the energy function to minimize the energy of states corresponding to the observed data. This can be expressed as minimizing the log-likelihood (Eq. 5). Since $\log Z(\theta)$ is intractable, SGLD is used to approximate the gradient of the log-likelihood w.r.t. the model parameters (Eq. 10).

In this paper, we leverage EBMs within the concept encoder to probabilistically infer the relationship between qCAVs and the variational latent vector v_k. By performing probabilistic relation modeling, we can obtain robust concepts that are more resilient to the variabilities and complexities of real-world scenarios.

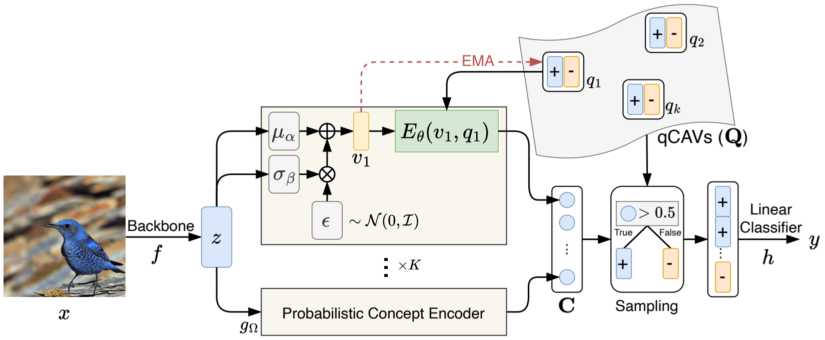


Fig. 2: Overall architecture of the EQ-CBM. The input image x is processed by the backbone network f to generate a latent vector z. This latent vector is fed into probabilistic concept encoders g_Ω, which use variational inference techniques to infer v_k. The energy function E_θ evaluates the compatibility of v_k with the qCAVs. Exponential Moving Average (EMA) updates each vector pair. The sampling module selects the concept vectors having lowest energy scores, which are then used for the final task.

3 EQ-CBM

We introduce EQ-CBM, a framework that enhances CBMs by utilizing probabilistic concept encoding with EBMs with qCAVs. Our approach addresses the limitations of previous CBMs by improving robustness and interpretability in complex decision-making tasks. As depicted in Fig. 2, the proposed method comprises three main components: qCAVs (**Q**), probabilistic concept encoders (g_Ω), and a sampling module.

In this section, we describe each component and its integration into the overall framework, highlighting the advantages of using energy-based modeling and vector quantization for concept representation.

3.1 Quantized Concept Activation Vectors

Vector quantization is a powerful technique widely used to discretize continuous vectors into a finite set of representative vectors, known as a codebook. Notable applications of vector quantization include VQ-VAE [36], which avoids posterior collapse and enables high-quality generative modeling. By mapping continuous data to discrete codebook vectors, vector quantization captures the underlying structure of the data by aligning representations with a finite set of learned codebook vectors. This alignment is particularly beneficial for interpretability and robustness, as it simplifies the analysis and manipulation of the model's internal representations.

In our approach, we propose qCAVs to enhance the interpretability and robustness of CBMs. This discretization helps in standardizing the learned concepts, facilitating more straightforward model behavior and enabling effective human intervention. Moreover, quantized concepts provide a stable and consistent basis for decision-making, improving the model's reliability in complex real-world scenarios.

As shown in Fig. 2, qCAVs consist of K non-differentiable vector pairs: $\mathbf{Q} = \{(q_k^+, q_k^-) = q_k\}_{\in[K]}$, $q_k^+ \in \mathbb{R}^d$ and $q_k^- \in \mathbb{R}^d$. Each pair represents a positive and negative vector for a particular k^{th} concept. In our probabilistic concept encoder, these vectors are used as conditioning variables to define a joint energy-concept landscape. The detailed process for updating qCAVs is described in the following section.

3.2 Probabilistic Concept Encoder

The probabilistic concept encoders g_Ω abstract each concept by applying variational inference techniques to extract a variational latent vector v_k from a normal distribution. Specifically, as depicted in Fig. 2, from z, we infer the mean μ_α and variance σ_β, and sample noise ϵ from Gaussian distribution to learn diverse representations. This process enables the model to capture the variability and complexity of real-world concepts more effectively.

Our energy function E_θ then models the joint distribution between the variational latent vector v_k and the qCAVs. Instead of updating qCAVs through backpropagation gradients, we use an exponential moving average (EMA) to update each vector pair, which is more effective for ensuring stability and consistent updates. The EMA is applied as follows:

$$q_{k,s}^+ := \begin{cases} q_{k,s-1}^+ \cdot \eta + v_k \cdot (1-\eta) & c_k^* = 1 \\ q_{k,s-1}^+ & \textit{otherwise} \end{cases}, \tag{12}$$

$$q_{k,s}^- := \begin{cases} q_{k,s-1}^- \cdot \eta + v_k \cdot (1-\eta) & c_k^* = 0 \\ q_{k,s-1}^- & \textit{otherwise} \end{cases}, \tag{13}$$

where s represents the training steps and η is the decay factor, set to 0.95, ensuring a balance between the historical values of the qCAVs and the newly observed variational latent vectors v_k. This balance allows the model to adapt gradually without abrupt changes.

3.3 Energy-based Concept Encoding

To effectively capture the relationships between variational latent vectors v_k and qCAVs, we employ an energy-based approach for concept encoding. EBMs offer a robust and flexible framework for modeling complex dependencies and probabilistic relationships. By utilizing an energy function E_θ, we measure the compatibility between v_k and predefined qCAVs, enabling a more comprehensible and interpretable representation of concepts.

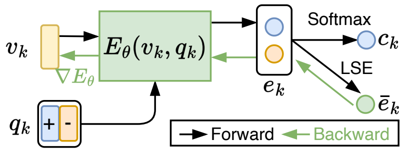


Fig. 3: Integration of an EBM in the concept encoder. The output is processed through softmax for concept prediction c_k and LogSumExp (LSE) for composed energy score $\bar{e}_k$. The backward pass using SGLD updates v_k based on the joint energy-concept landscape.

The core idea behind energy-based concept encoding is to define a joint energy-concept landscape where low energy scores indicate high compatibility between v_k and the qCAVs. This approach allows the model to probabilistically infer concept activations given v_k, thereby effectively capturing the variability and complexity of real-world scenarios. As shown in Fig. 3, we integrate an EBM into the concept encoders, which operates on the variational latent vector v_k. This energy function returns a low value when v_k is closely related to the q_k, facilitating the selection of the most relevant concept vectors for the final task. The problem we aim to solve is therefore represented as the joint distribution of c_k, v_k, and q_k as follows:

$$\log p_\Omega(c_k, v_k, q_k) = \log p_\theta(v_k, q_k) + \log p_\Omega(c_k|v_k, q_k), \tag{14}$$

where $p_\Omega(c_k|v_k, q_k)$ is normalized w.r.t. c_k, making it straightforward to compute by maximizing the likelihood. However, since EBMs are unnormalized, directly maximizing the likelihood is more challenging. Therefore, we estimate the gradient of the following likelihood:

$$\begin{aligned}\nabla_\Omega \mathbb{E}_{p_\mathcal{D}(c_k, v_k, q_k)}[\log p_\Omega(c_k, v_k, q_k)] &= \nabla_\Omega \mathbb{E}_{p_\mathcal{D}(c_k, v_k, q_k)}[\log p_\Omega(c_k|v_k, q_k)] \\ &+ \nabla_\theta \mathbb{E}_{p_\theta(v'_k, q_k)}[\nabla_\theta E_\theta(v'_k, q_k)] - \nabla_\theta \mathbb{E}_{p_\mathcal{D}(v_k, q_k)}[\nabla_\theta E_\theta(v_k, q_k)],\end{aligned} \tag{15}$$

where the first term is equivalent to cross-entropy and can be treated similarly to a standard classification task conditioned on q_k. The second term, $\nabla_\theta \mathbb{E}_{p_\theta(v'_k, q_k)}[\nabla_\theta E_\theta(v'_k, q_k)]$, forms gradients that increase the energy function $E_\theta(v'_k, q_k)$ for the negative samples v'_k. Conversely, the third term reduces the energy for the actual samples v_k, learning to minimize the energy for these real samples.

Following previous EBMs [21,11,16], we use SGLD to synthesize negative samples for the second term in Eq. 11. SGLD generates samples that approximate the true distribution. To facilitate this, we redefine the problem as a binary classification task to determine whether the variational latent vector v_k is related to either q_k^+ or q_k^-. Consequently, the energy function is defined as $E_\theta : (v_k, q_k) \rightarrow e_k \in \mathbb{R}^2$. Unlike prior approaches that utilized a single scalar value for energy scores, we interpret the energy function as class logits. By employing the LogSumExp (LSE) function, as shown in Eq. 16, the model concurrently addresses both classification and energy modeling, inspired by the work of Grathwohl *et al.* [8]. The update rule for SGLD is as follows. For simplicity, we omit the indexing k:

$$v_0 \sim p_0(v), \quad v_{t+1} = v_t - \frac{\gamma}{2} \frac{\partial \log \sum_c \exp(E_\theta(v_t, q)[c])}{\partial v_t} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \gamma) \tag{16}$$

Algorithm 1: Learning algorithm

```
Input: image $x$, ground truth concepts $\mathbf{C}^* = \{c_1^*, \dots, c_k^*\}$, qCAVs $\mathbf{Q} = \{q_1, \dots, q_k\}$
$\mathbf{C} \leftarrow \emptyset$
$\bar{e}$, $\bar{e}' \leftarrow 0$
$z \leftarrow f(x)$                                        /* Backbone network */
foreach $\{\alpha, \beta, \theta\}^K \subseteq \Omega$ do
    $v_k \leftarrow \mu_\alpha(z) + \sigma_\beta(z) \cdot \epsilon$                    /* Variational inference */
    $e_k \leftarrow E_\theta(v_k, q_k)$
    $\bar{e} \leftarrow \bar{e} + \mathrm{LSE}(e_k)$
    $v'_k \leftarrow \mathcal{N}(0, \mathcal{I})$
    for $t = 1$ to $T$ do
        $v'_k \leftarrow \mathrm{SGLD}(v'_k, q_k, c_k^*)$                           /* Eq. 16 */
    end
    $e'_k \leftarrow E_\theta(v'_k, q_k)$
    $\bar{e}' \leftarrow \bar{e}' + \mathrm{LSE}(e'_k)$
    $\mathbf{C} \leftarrow \mathbf{C} \cup \mathrm{Softmax}(e_k)[+]$          /* Obtain the concept score for $q_k^+$ */
    $q_k \leftarrow \mathrm{EMA}(v_k, q_k, c_k^*)$                                /* Eqs. 12-13 */
end
$y \leftarrow h(\mathrm{sampling}(\mathbf{Q}, \mathbf{C}))$
```

To train the energy function, we apply contrastive divergence loss, as represented in Eq. 15, which offers several benefits. Firstly, it enables the model to learn robust representations by contrasting the energy of positive samples against negative samples. Secondly, it regularizes the model by preventing the energy values from diverging. The energy-based objective function is defined as follows:

$$\mathcal{L}_e = \frac{1}{K} \sum_k \left(\bar{e}'_k - \bar{e}_k + (\bar{e}'_k + \bar{e}_k)^2 \right) \tag{17}$$

$$\bar{e}'_k = -\mathrm{LSE}\left(E_\theta(v'_k, q_k) \right), \quad \bar{e}_k = -\mathrm{LSE}\left(E_\theta(v_k, q_k) \right), \tag{18}$$

The overall learning process for the proposed method is summarized in Algorithm 1, with the final objective function for training the proposed model defined as follows:

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_{\mathrm{CE}}(\mathbf{C}, \mathbf{C}^*) + \lambda_y \mathcal{L}_{\mathrm{CE}}(y, y^*) + \lambda_e \mathcal{L}_e \tag{19}$$

4 Experiments

To evaluate our model, we used four datasets: CUB [38], CelebA [22], AwA2 [39], and TravelingBirds [20]. The CUB dataset includes 12K bird images across 200 species, with 6K for training and 6K for testing, each labeled with 312 attributes. Following previous works [20,6,17], we used 112 attributes as concepts. CelebA contains 202K facial images of 10K celebrities, each with 40 attributes, from which we used six attributes as concepts. The AwA2 dataset has 37K images

Table 1: Hyperparameters.

Datasets	d	T	γ	λ_c	λ_y	λ_e	LR
CUB	16	20	0.4	5	1	0.05	0.005
CelebA	16	20	0.4	5	1	0.05	0.01
AwA2	16	20	0.4	5	1	0.05	0.005

Table 2: Comparison of concept and task accuracy across various CBM models (**Bold**: best score, underline: second-best score).

Methods	CUB [37]		CelebA [22]		AwA2 [39]	
	Concept (±CI)	Task (±CI)	Concept (±CI)	Task (±CI)	Concept (±CI)	Task (±CI)
Fuzzy-CBM [20]	95.882 (±0.105)	74.228 (±0.606)	90.269 (±0.211)	33.696 (±2.104)	99.000 (±0.168)	95.089 (±1.005)
Bool-CBM [20]	96.229 (±0.031)	72.512 (±0.466)	90.329 (±0.164)	33.915 (±0.885)	99.001 (±0.188)	94.869 (±1.047)
CEM [6]	96.160 (±0.157)	79.029 (±0.519)	90.237 (±0.306)	42.618 (±1.412)	99.048 (±0.037)	95.745 (±0.294)
Prob-CBM [17]	95.596 (±0.061)	76.265 (±0.145)	89.272 (±0.238)	34.472 (±0.893)	98.283 (±0.065)	92.485 (±0.315)
Coop-CBM [34]	89.892 (±0.649)	79.154 (±0.734)	90.534 (±0.142)	42.393 (±1.354)	98.875 (±0.107)	95.927 (±0.153)
ECBM [40]	96.536 (±0.091)	77.148 (±0.695)	90.006 (±0.986)	34.976 (±2.111)	98.909 (±0.037)	94.555 (±0.121)
EQ-CBM (Ours)	**96.580 (±0.043)**	**79.310 (±0.272)**	**90.617 (±0.309)**	**56.600 (±1.060)**	**99.129 (±0.022)**	**95.965 (±0.102)**

of 50 animal species, annotated with 85 attributes. The TravelingBirds dataset, derived from the CUB dataset [38], replaces image backgrounds with diverse scenes [44] to test model robustness in real-world uncertainties. All images were resized to 299×299 pixels across datasets. All experiments were conducted on a system with an AMD 5955WX CPU and an Nvidia A6000 GPU using five random seeds. Detailed hyperparameters are depicted in Table 1.

4.1 Metrics

To comprehensively evaluate the performance, we utilized several key metrics: Concept Accuracy, Task Accuracy, and Uncertainty. Each metric provides insight into different aspects of model performance, from the precision of concept predictions to the robustness of the model under uncertain conditions.

Concept Accuracy measures how accurately the model predicts predefined concepts. High concept accuracy indicates effective learning and prediction of concepts.

Task Accuracy measures the correctness of the model's primary classification task. High task accuracy means the model performs well in predicting the correct class using only concept.

Uncertainty assesses the model's confidence and robustness to variability in the data. It is calculated as follows:

$$e_k = (e_k^+, e_k^-) = E_\theta(v_k, q_k), \quad u_k = \big(\exp\big(\mathrm{LSE}(e_k) - \mathrm{Mean}(e_k)\big) - 1\big)^{-1}, \tag{20}$$

where (e_k^+, e_k^-) are the paired energy scores for a concept k, and u_k represents the uncertainty, with higher values indicating greater uncertainty.

4.2 Concept and Task accuracy

To evaluate the effectiveness of our proposed model, we compared its performance against several CBMs using three datasets: CUB, CelebA, and AwA2. We measured both concept prediction accuracy and task (classification) accuracy. Table 2 presents the results of these experiments, showing the mean accuracy and confidence intervals (CI) for each method.

As shown in Table 2, our proposed EQ-CBM outperforms existing approaches in both concept and task accuracy across all three datasets. For the CUB

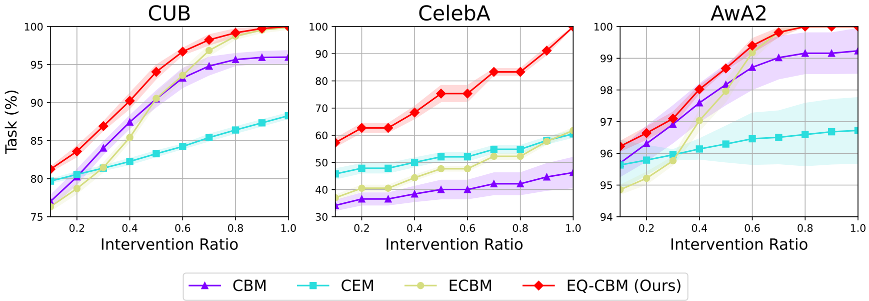


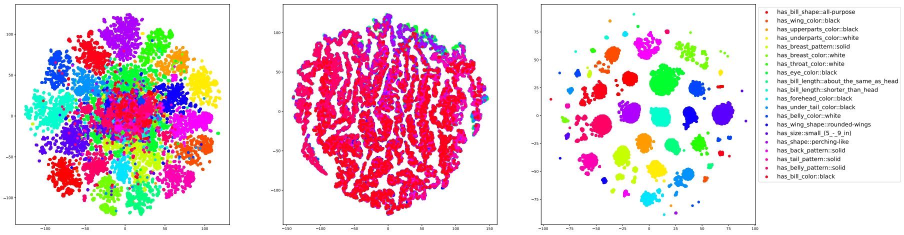
Fig. 4: Task accuracy under varying levels of human intervention in concept predictions for different models.

dataset, EQ-CBM achieved a concept accuracy of 96.580% and a task accuracy of 79.310%. In comparison, the best-performing baseline, ECBM, achieved a concept accuracy of 96.536% but a lower task accuracy of 77.148%. Similarly, for the CelebA dataset, our model attained concept and task accuracies of 90.617% and 56.600%, respectively, significantly surpassing the best-performing baseline, CEM, which had a task accuracy of 42.618%. In the AwA2 dataset, EQ-CBM reached concept and task accuracies of 99.129% and 95.965%, respectively. The closest competitor, Coop-CBM, achieved 98.875% in concept accuracy and 95.927% in task accuracy.

These results demonstrate that our model not only maintains high concept accuracy but also achieves superior task performance, validating the effectiveness of incorporating EBMs with qCAVs. Notably, our approach consistently outperformed other models in task accuracy, particularly on the CelebA dataset, where it exceeded the second-best method by a substantial margin (56.600% vs. 42.618%). This highlights the robustness and generalization capability of our model across different datasets and tasks. The significant improvements in task accuracy, especially on challenging datasets such as the CelebA dataset, underscore the effectiveness of qCAVs in capturing complex relationships between concepts and improving overall model performance. These findings suggest that EQ-CBM can be a valuable addition to existing concept-based models, providing both enhanced interpretability and higher task accuracy.

4.3 Concept Intervention

To further evaluate the robustness and interpretability of our model, we conducted a series of concept intervention experiments. The goal of these experiments is to assess how human intervention in correcting concept predictions influences the overall task accuracy. As illustrated in Fig. 4, we compared the performance of our model (EQ-CBM) with several methods, including CBM, CEM, and ECBM, under varying levels of intervention. The intervention ratio on the x-axis represents the proportion of corrected concept predictions, while the y-axis shows the corresponding task accuracy.



(a) CEM [6] **(b)** ECBM [40] **(c)** EQ-CBM (Ours)

Fig. 5: t-SNE visualizations of encoded concepts for (a) CEM, (b) ECBM, and (c) EQ-CBM (Ours). Different colors represent different concepts.

For the CUB dataset, as the intervention ratio increased, EQ-CBM consistently outperformed other models, achieving nearly 100% task accuracy at higher intervention ratios. This demonstrates the effectiveness of our model in leveraging homogeneous concepts from qCAVs to enhance human intervention. Similarly, in the CelebA dataset, our model showed significant improvements in task accuracy with increasing intervention ratios, outperforming all baseline methods. Notably, EQ-CBM achieved a task accuracy of 100% with full intervention, highlighting its robustness and capacity to incorporate human feedback effectively. In the AwA2 dataset, EQ-CBM again outperformed the baseline methods across all intervention levels. With an intervention ratio of 1.0, our model reached a task accuracy of 100%, indicating its superior ability to utilize qCAVs for improved the level of human intervention.

These intervention experiments underscore the advantages of our approach in terms of both robustness and interpretability. By allowing for human intervention in concept predictions, EQ-CBM can significantly enhance the overall accuracy of the model. This is particularly beneficial in real-world applications where model predictions can be iteratively refined through human expertise.

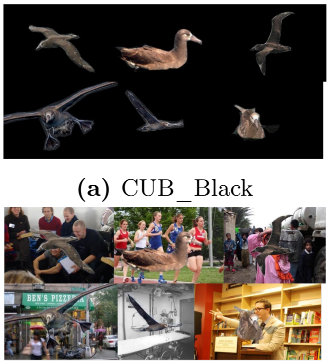
4.4 Visualization of Encoded Concepts

To assess how well the encoded concepts capture the underlying data structure, we utilized t-SNE to visualize the concepts produced by each model just before the final task linear classifier. Figure 5 shows the t-SNE plots of these concepts for CEM [6], ECBM [40], and our proposed EQ-CBM on the CUB dataset.

In Fig. 5a, the encoded concepts produced by CEM exhibited significant overlap among different concepts, indicating weak separation and potential confusion in concept interpretation. This overlap suggests that the model might struggle to differentiate between certain concepts, leading to less accurate and interpretable predictions. In Fig. 5b, while ECBM achieves better separation than CEM, there is still considerable clustering within certain concept groups. This partial overlap could impact the model's ability to clearly distinguish between similar concepts, potentially affecting its task performance and interpretability. In contrast, our proposed EQ-CBM, as depicted in Fig. 5c, demonstrated a much clearer separation of concepts. The t-SNE plot reveals distinct clusters for each concept, indicating that EQ-CBM effectively captures the unique characteristics of each concept and represents them in a well-separated latent space. Specifically, these concepts are sampled qCAVs, which provide a homogeneous representation of concepts by incorporating variability and uncertainty. This clear separation enhances the model's interpretability and allows for more accurate concept-based predictions.

4.5 Uncertainty Robustness

To evaluate the robustness of our proposed model under uncertain conditions, we conducted experiments using the TravelingBirds dataset, a variant of the CUB dataset where bird image backgrounds are replaced with various real-world scenes from the Places dataset [44]. This introduced background variability, challenging the model's ability to maintain performance under these uncertain conditions. The models were trained on the CUB dataset and tested on two versions of the TravelingBirds dataset: CUB_Black, where backgrounds are replaced with black (Fig. 6a), and CUB_Random, where backgrounds are replaced with random real-world scenes (Fig. 6b).



(a) CUB_Black

(b) CUB_Random

Fig. 6: Randomly selected samples from the TravelingBirds dataset.

Figure 7 presents the results of our uncertainty robustness experiments, comparing the task accuracy of our EQ-CBM with several baseline methods. In the CUB_Black scenario, our EQ-CBM achieved the highest task accuracy, outperforming other methods. ProbCBM performed comparably but slightly lower than ours, while other methods showed reduced accuracy, indicating that our approach is more robust to background variations. Similarly, in the CUB_Random scenario, EQ-CBM again outperformed all baseline methods. Coop was the closest competitor, but still showed slightly lower task accuracy compared to ours. Other models exhibited even lower performance, reinforcing the robustness of our model in handling diverse and unpredictable backgrounds.

Overall, the experimental results indicated that ProbCBM and ECBM are more influenced by background variations (*i.e.*, habitat) than by object-centric concepts, whereas our EQ-CBM demonstrated superior robustness to these variations.

4.6 Ablation Study

We performed an ablation study to evaluate the impact of different components of our model on both concept and task accuracy. The results are shown in Table 3. Without EMA, the model's concept accuracy dropped to 92.081% and task accuracy significantly decreased to 23.745%. This indicates that EMA is

Table 3: Impact of model components on concept and task accuracy on the CUB dataset.

	Concept (±CI)	Task (±CI)
w/o EMA	92.081 (±0.633)	23.745 (±3.915)
w/o Energy	95.454 (±0.167)	78.256 (±0.463)
w/o Var. infer.	96.518 (±0.063)	78.681 (±0.433)
EQ-CBM	**96.580 (±0.043)**	**79.310 (±0.272)**

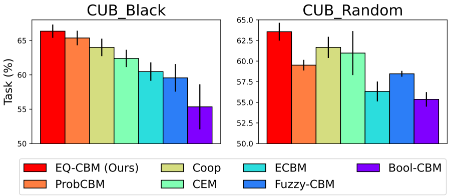


Fig. 7: Task accuracy on the Traveling-Birds dataset.

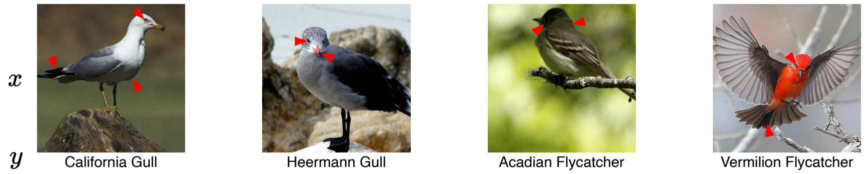
crucial for stabilizing and ensuring consistent convergence of qCAVs. Removing the energy-based modeling component led to a concept accuracy of 95.454%, demonstrating that energy-based modeling is essential for capturing nuanced relationships between concepts, resulting in higher accuracy. Training without variational inference showed a slight decrease in task accuracy to 78.681%, with concept accuracy at 96.518%. This suggests that variational inference helps in learning diverse features, slightly enhancing performance. In contrast, the complete EQ-CBM achieved the highest performance, with a concept accuracy of 96.580% and a task accuracy of 79.310%, confirming the effectiveness of combining EMA, energy-based modeling, and variational inference.

4.7 Concept Interpretation

To demonstrate the interpretability of our model, we visualized several concepts and their corresponding scores and uncertainties on the CUB dataset, as shown in Fig. 8. Each image depicts a bird species with its corresponding concepts, predicted concept scores c_k, and uncertainties u_k.

For the California Gull, the model incorrectly identifies the `belly_color` due to shadows, resulting in high uncertainty. The `upper_tail_color` also shows high uncertainty because it is occluded. The `eye_color` is correctly identified but with high uncertainty due to its multicolored nature. For the Heermann Gull, the `eye_color` and `bill_color` are correctly identified but have high uncertainty because of the small size of the eyes and the multicolored bill. In the Acadian Flycatcher image, the `nape_color` prediction has some uncertainty due to a shadow that makes the nape appear gray. Although the predicted concept is incorrect, the uncertainty score indicates that the model is unsure about this prediction. Similarly, the `eye_color` prediction also shows high uncertainty due to the shadow effect. For the Vermilion Flycatcher, the model predicts the `under_tail_color` with some uncertainty. While the `under_tail` is not entirely black, the presence of some black in the actual tail leads to relatively high uncertainty. The `eye_color` prediction also shows high uncertainty because the bird is facing forward, making it difficult for the model to judge accurately.

These visualizations help in understanding the model's decision-making process, highlighting where the model is less confident. This interpretability is crucial for validating the model's predictions and identifying areas for potential improvement.



	c_k	u_k
has_belly_color::white	0.44	0.99
has_upper_tail_color::white	0.44	0.99
has_eye_color::balck	0.63	0.93
has_bill_shape::dagger	0.03	0.01

	c_k	u_k
has_eye_color::balck	0.61	0.95
has_bill_color::black	0.36	0.93
has_belly_color::white	0.00	0.05
has_leg_color::black	0.94	0.00

	c_k	u_k
has_nape_color::grey	0.54	0.99
has_eye_color::black	0.65	0.91
has_shape::perching-like	1.00	0.02
has_leg_color::grey	0.00	0.01

	c_k	u_k
has_eye_color::black	0.55	0.99
has_under_tail_color::black	0.36	0.92
has_bill_shape::cone	0.20	0.68
has_leg_color::black	0.99	0.05

Fig. 8: Interpretation of several concepts. Green boxes indicate true predictions, while red boxes indicate false predictions. c_k is the concept score and u_k is the uncertainty.

5 Conclusion

In this paper, we introduced EQ-CBM, a novel framework designed to enhance CBMs through probabilistic concept encoding using EBMs with qCAVs. Our approach addressed the limitations of deterministic concept encoding in existing CBMs by enabling robust probabilistic inference, thereby improving both concept and task accuracy. By integrating EBMs, our framework effectively captured the underlying dependencies and uncertainties within the data, leading to more accurate and interpretable models. qCAVs ensured the selection of homogeneous vectors during concept encoding, enhancing both task performance and human intervention capabilities. Experiments on datasets such as CUB, AwA2, CelebA, and TravelingBirds demonstrated the superiority of EQ-CBM in achieving a better balance between interpretability and accuracy compared to existing CBMs. The TravelingBirds dataset further showcased the robustness of our model under challenging conditions. Our ablation studies underscored the importance of each component within our framework.

Future work will focus on improving the scalability of our approach to efficiently handle larger datasets, providing a robust and flexible framework for enhancing the decision-making processes of DNNs.

Acknowledgments. This research was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [24ZD1120, Development of ICT Convergence Technology for Daegu-GyeongBuk Regional Industry].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for boltzmann machines. Cognitive science **9**(1), 147–169 (1985)

2. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV. pp. 839–847 (2018)
3. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., Dvijotham, K.: Interactive concept bottleneck models. In: AAAI. pp. 5948–5955 (2023)
4. Du, Y., Li, S., Sharma, Y., Tenenbaum, J., Mordatch, I.: Unsupervised learning of compositional energy concepts. In: NeurIPS. pp. 15608–15620 (2021)
5. Du, Y., Mordatch, I.: Implicit generation and modeling with energy based models. In: NeurIPS. pp. 1–11 (2019)
6. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., Jamnik, M.: Concept embedding models: Beyond the accuracy-explainability trade-off. In: NeurIPS. pp. 21400–21413 (2022)
7. Gao, R., Lu, Y., Zhou, J., Zhu, S.C., Wu, Y.N.: Learning generative convnets via multi-grid modeling and sampling. In: CVPR. pp. 9155–9164 (2018)
8. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. In: ICLR. pp. 1–23 (2020)
9. Guo, Q., Ma, C., Jiang, Y., Yuan, Z., Yu, Y., Luo, P.: Egc: Image generation and classification via a diffusion energy-based model. In: ICCV. pp. 22952–22962 (2023)
10. Han, T., Lu, Y., Zhu, S.C., Wu, Y.N.: Alternating back-propagation for generator network. In: AAAI. pp. 1–10 (2017)
11. Han, T., Nijkamp, E., Zhou, L., Pang, B., Zhu, S.C., Wu, Y.N.: Joint training of variational auto-encoder and latent energy-based model. In: CVPR. pp. 7978–7987 (2020)
12. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing leakage in concept bottleneck models. In: NeurIPS. pp. 23386–23397 (2022)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation **18**(7), 1527–1554 (2006)
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: ICML. pp. 2668–2677 (2018)
16. Kim, B., Ye, J.C.: Energy-based contrastive learning of visual representations. In: NeurIPS. pp. 4358–4369 (2022)
17. Kim, E., Jung, D., Park, S., Kim, S., Yoon, S.: Probabilistic concept bottleneck models. In: ICML. pp. 16521–16540 (2023)
18. Kim, S., Ko, B.C.: Concept Graph Embedding Models for Enhanced Accuracy and Interpretability. Machine Learning: Science and Technology **5**(3), 1–15 (2024)
19. Kim, S., Nam, J., Ko, B.C.: ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder. In: ICML. pp. 11162–11172 (2022)
20. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML. pp. 5338–5348 (2020)
21. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1** (2006)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV. pp. 3730–3738 (2015)
23. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS. pp. 4765–4774 (2017)

24. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: CVPR. pp. 14933–14943 (2021)
25. Neal, R.M.: Mcmc using hamiltonian dynamics. Handbook of markov chain monte carlo **2**(11), 2 (2011)
26. Nijkamp, E., Hill, M., Han, T., Zhu, S.C., Wu, Y.N.: On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In: AAAI. pp. 5272–5280 (2020)
27. Nijkamp, E., Hill, M., Zhu, S.C., Wu, Y.N.: Learning non-convergent non-persistent short-run mcmc toward energy-based model. In: NeurIPS. pp. 361–378 (2019)
28. Nijkamp, E., Pang, B., Han, T., Zhou, L., Zhu, S.C., Wu, Y.N.: Learning multi-layer latent variable model via variational optimization of short run mcmc for approximate inference. In: ECCV. pp. 361–378 (2020)
29. Pang, B., Han, T., Nijkamp, E., Zhu, S.C., Wu, Y.N.: Learning latent space energy-based prior model. In: NeurIPS. pp. 21994–22008 (2020)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)
31. Salakhutdinov, R., Hinton, G.: Deep boltzmann machines. In: Artificial intelligence and statistics. pp. 448–455 (2009)
32. Sarkar, A., Vijaykeerthy, D., Sarkar, A., Balasubramanian, V.N.: A framework for learning ante-hoc explainable models via concepts. In: CVPR. pp. 10286–10295 (2022)
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
34. Sheth, I., Ebrahimi Kahou, S.: Auxiliary losses for learning generalizable concept-based models. In: NeurIPS. pp. 1–25 (2024)
35. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: ICML. pp. 3145–3153 (2017)
36. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS. pp. 1–10 (2017)
37. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-ucsd birds-200-2011 dataset. Computation & Neural Systems Technical Report, CNS-TR-2011-001 (2011), http://www.vision.caltech.edu/visipedia/CUB-200-2011.html
38. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Tech. rep. (2010)
39. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2251–2265 (2018)
40. Xu, X., Qin, Y., Mi, L., Wang, H., Li, X.: Energy-based concept bottleneck models. In: ICLR. pp. 1–23 (2023)
41. Yang, X., Ji, S.: Jem++: improved techniques for training jem. In: ICCV. pp. 6494–6503 (2021)
42. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: ICLR. pp. 1–20 (2023)
43. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. In: ICLR. pp. 1–17 (2017)
44. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(6), 1452–1464 (2017)

45. Zhu, S.C., Mumford, D.: Grade: Gibbs reaction and diffusion equations. In: ICCV. pp. 847–854 (1998)