CoCo-Bot: Energy-based Composable Concept Bottlenecks for Interpretable Generative Models

Sangwon Kim In-Su Jang Pyong-Kun Kim Kwang-Ju Kim*
ETRI
{eddiekim, jef1015, iros, kwangju}@etri.re.kr

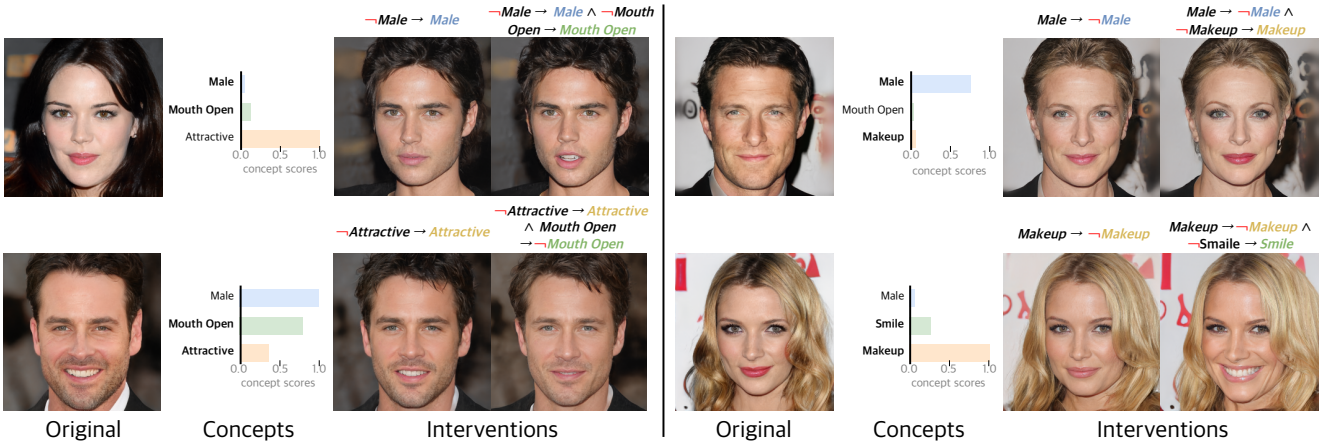

Figure 1. **Interpretable and composable concept-based interventions with CoCo-Bot.** The "Original" column shows images generated from latent vectors without any concept-level intervention. The "Concepts" column presents predicted scores for each concept in the corresponding samples. The "Interventions" column illustrates how composing (∧) and negating (¬) specific concepts alters the output through direct user control. These changes are entirely driven by explicit concept intervention on the energy landscape, enabling transparent and interpretable generative control.

Abstract

Concept Bottleneck Models (CBMs) were originally designed to enhance interpretability in classification by enforcing predictions through human-understandable concepts. Their recent extension to generative models offers semantic-level control via concept interventions. However, existing generative CBMs often depend on auxiliary visual cues at the bottleneck, which can compromise interpretability and compositionality. We propose ***CoCo-Bot****, a post-hoc composable concept bottleneck for generative models that eliminates auxiliary cues and ensures that all semantic control is achieved via explicit concepts. Diffusion-based energy functions enable robust post-hoc interventions, including concept composition and negation. Experiments using StyleGAN2 pre-trained on CelebA-HQ show that CoCo-Bot improves concept-level controllability and interpretability, while maintaining competitive visual quality.*

1. Introduction

Concept Bottleneck Models (CBMs) [1, 6, 7, 12] were originally proposed to enhance interpretability in neural networks by introducing intermediate predictions over explicit, human-understandable concepts. This paradigm has recently been extended to deep generative models [3, 8], enabling semantic-level interpretation and post-hoc interventions in generative processes. However, applying CBMs to generative tasks introduces unique challenges, as the model must reconstruct high-dimensional outputs from a limited set of semantic concepts—often at the cost of expressiveness.

To address this, prior work [8] introduced auxiliary visual cues at the bottleneck. These additional cues help capture information beyond the provided concepts but also compromise transparency. When the model depends on latent, unobserved features, it becomes difficult to ensure that concept-level interventions produce clear and reliable changes in the output.

We propose **CoCo-Bot** (**Co**mposable **Co**ncept **Bot**tleneck Generative Model), an energy-based post-hoc CBM framework that removes all auxiliary vision cues from the bottleneck. CoCo-Bot constrains the generative process to operate solely through explicit, compositional concepts, ensuring that any intervention is directly and faithfully reflected in the generated samples. This design enables robust, user-controllable semantic intervention—such as concept composition and negation—while preserving interpretability.

We evaluate CoCo-Bot using StyleGAN2 [5] pre-trained on the CelebA-HQ [4] dataset. Experimental results demonstrate that, compared to prior CBM-based generative models, CoCo-Bot achieves substantially improved compositional control and transparency, while maintaining competitive generative quality.

2. Related Works

Concept Bottleneck Models and Generative Extensions: CBMs [1, 6, 7, 12] were originally introduced to enhance interpretability in classification models by requiring intermediate predictions over human-defined concepts. Subsequent works have extended this framework to embedding spaces [1] and, more recently, to deep generative models [3, 8], with the goal of enabling semantic-level control and post-hoc intervention. However, some generative CBM approaches, such as A. Kulkarni *et al*. [8], introduce auxiliary vision cues or additional latent dimensions at the bottleneck to capture information not explained by the provided concepts. This reliance on unobserved channels can limit interpretability and weaken compositionality, making it challenging to ensure that interventions on multiple concepts are consistently and transparently reflected in the generated outputs.

Energy-based Models for Concept Bottlenecks: Energy-based models (EBMs) [2, 11] have re-emerged as a flexible class of generative models, notable for representing unnormalized densities and supporting control through gradient-based guidance [9]. While recent works have combined EBMs with concept bottlenecks [6, 12], most prior EBM-based CBMs address classification or deterministic prediction, not fully generative modeling with complex outputs. Conventional EBMs typically require Markov Chain Monte Carlo (MCMC) methods such as Stochastic Gradient Langevin Dynamics (SGLD) for inference and sampling, due to the intractability of the partition function Z. However, these MCMC-based approaches are computationally intensive and often unstable for high-dimensional data. This limitation poses a barrier to developing energy-based generative CBMs that can efficiently and reliably support interpretable concept interventions.

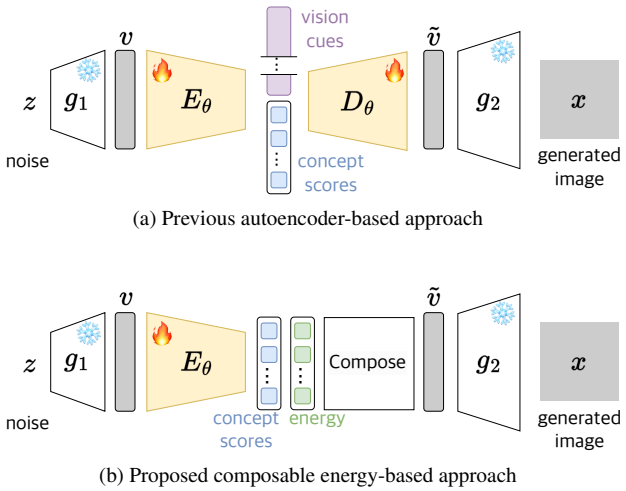

Figure 2. **Comparison of concept bottlenecks in generative models.** (a) Prior work uses auxiliary vision cues at the bottleneck. (b) CoCo-Bot enforces generation only through explicit, composable concepts.

3. CoCo-Bot

Figure 2 highlights the key distinction between prior autoencoder-based CBMs and our proposed CoCo-Bot. In the previous method, auxiliary vision cues at the bottleneck absorb residual information not captured by explicit concepts, which limits both interpretability and compositional intervention. In contrast, CoCo-Bot eliminates this auxiliary pathway entirely, ensuring that generative representation flows solely through human-interpretable concepts.

This explicit design compels the model to encode all representation relevant to generation in a set of semantic concepts. As a result, any user intervention—such as composing or negating concepts—yields predictable and transparent changes in the output. Users can reliably perform a wide range of semantic edits simply by adjusting concept assignments, without retraining or additional supervision.

3.1. Training

CoCo-Bot is trained as an energy-based model (E_θ) that reconstructs images conditioned on provided concept vectors (c_k), thereby enforcing interpretability via explicit compositional control.

Data Preparation: As shown in Fig. 2, we first sample a latent vector v from the StyleGAN2 [5] mapping network (g_1) given random noise z. Passing v through the synthesis network (g_2) produces an image x. Following A. Kulkarni *et al*. [8], a pseudo-labeler pre-trained on CelebA-HQ [4] infers K concept pseudo-labels for x, enabling self-supervised concept-level supervision.

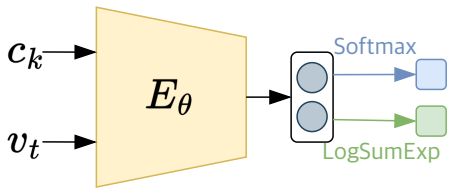

Figure 3. **CoCo-Bot module.** The model produces per-concept logits, applies a Softmax for concept prediction, and aggregates them via LogSumExp to yield a per-concept energy. This design supports both classification supervision and energy-based inference.

Diffusion-based Noising Process: Rather than relying on unstable and computationally expensive MCMC-based sampling, we employ a diffusion-based scheduler [10] to efficiently sample from the energy landscape. At each diffusion timestep t, a noisy latent v_t is constructed from the clean latent v as follows:

$$v_t = \sqrt{\alpha_t} v_0 + \sqrt{1-\alpha_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I) \tag{1}$$

where v_0 is the clean latent vector, α_t is a noise schedule, and ϵ_t is Gaussian noise.

Energy Function Optimization: The probability of a latent sample v is defined in terms of a composed energy:

$$p_\theta(v) = \frac{\exp(-\mathcal{E}_\theta(v))}{Z(\theta)}, \tag{2}$$

where $\mathcal{E}_\theta(v)$ is the total energy and $Z(\theta)$ is the intractable partition function. In CoCo-Bot, the total energy is the sum of per-concept energies:

$$\mathcal{E}_\theta(v_t; \mathbf{C}, t) = \sum_{k=1}^{K} e_\theta(v_t, c_k, t), \tag{3}$$

where each per-concept energy is defined as

$$e_\theta(v_t, c_k, t) = \text{LogSumExp}(E_\theta(v_t, c_k, t)), \tag{4}$$

with $E_\theta(\cdot)$ denoting the per-concept logits (see Fig. 3 for the architectural details). LogSumExp serves as a smooth approximation of the maximum logit, enabling compatibility with standard classification losses and facilitating joint training with per-concept classifiers.

Training Objective: The overall loss combines diffusion-based score matching with concept supervision:

1) Score matching. For a randomly sampled timestep t, the model minimizes the diffusion score-matching loss:

$$\mathcal{L}_{\text{score}} = \mathbb{E}_q \left[\frac{1}{2} \left\| \epsilon - \nabla_{v_t} \mathcal{E}_\theta(v_t; \mathbf{C}, t) \right\|^2 \right], \tag{5}$$

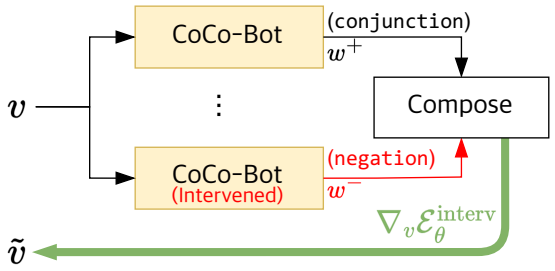

Figure 4. **Compositional concept interventions.** CoCo-Bot enables explicit composition and negation of user-specified concepts. Each intervention produces direct and predictable changes in the generated content, supporting transparent and interpretable control.

where ϵ is the noise added during the forward diffusion process, and $\nabla_{v_t}\mathcal{E}_\theta$ represents the gradient of the composed energy with respect to v_t, which acts as the model's score function.

2) Concept supervision. For each concept, the logits are supervised with a cross-entropy loss using the pseudo-labels:

$$\mathcal{L}_{\text{concept}} = -\sum_{k=1}^{K} \log \text{softmax}(E_\theta(v_t, c_k, t))[\hat{c}_k], \tag{6}$$

where $\hat{c}_k$ is the pseudo-label for concept k.

The final training loss is then:

$$\mathcal{L} = \mathcal{L}_{\text{score}} + \gamma \mathcal{L}_{\text{concept}}, \tag{7}$$

where γ balances concept classification and energy-based guidance.

3.2. Inference

At inference time, CoCo-Bot provides direct, interpretable intervention over generation by enabling explicit user interventions on concepts (see Fig. 4). This is achieved by introducing a weighted compositional energy:

$$\mathcal{E}_\theta^{\text{interv}}(v_t; \mathbf{C}, \mathbf{w}, t) = \sum_{k=1}^{K} w_k \, e_\theta(v_t, c_k, t), \tag{8}$$

where $w_k \in \{w^+, w^-\}$ encodes the user's intervention for each concept ($w^+ = 1$ for activation, $w^- = -0.001$ for negation).

Starting from an initial noise vector $\tilde{v}_t \sim \mathcal{N}(0, I)$ and a specified set of concept interventions, we iteratively update the latent as follows:

$$\tilde{v}_{t-1} = \tilde{v}_t - \eta \nabla_{\tilde{v}_t} \mathcal{E}_\theta^{\text{interv}}(\tilde{v}_t; \mathbf{C}, \mathbf{w}, t), \tag{9}$$

where η denotes the step size. This gradient-based procedure ensures that, at each iteration, the generative process

Table 1. **Comparison of concept accuracy and FID between CC-AE [8] and CoCo-Bot on CelebA-HQ [4].** CoCo-Bot achieves higher concept accuracy and comparable FID compared to the previous post-hoc concept bottleneck baseline, demonstrating enhanced interpretability without sacrificing image quality.

Method	Concept Acc. (%, ↑)	FID (↓)
CC-AE [8]	74.38	9.77
CoCo-Bot (*Ours*)	**75.70**	**6.47**

is steered precisely and transparently by the user-specified concept interventions, enabling faithful and compositional semantic control over the output.

Because all generative interventions are mediated solely through explicit, composable concepts, CoCo-Bot provides robust interpretability, transparent compositionality, and faithful human-in-the-loop editing as well as counterfactual exploration.

4. Experiments

We empirically evaluate CoCo-Bot on the CelebA-HQ [4] dataset, following the experimental protocol of prior concept bottleneck generative models such as CC-AE [8]. All results are reported on a set of 5K samples randomly generated from the same latent seeds for both CC-AE [8] and our model, ensuring a fair and direct comparison. For all experiments, we use $K = 8$ semantic concepts as defined in prior work, and set the loss balancing coefficient $\gamma = 10^{-3}$.

Our experiments address two primary questions: (1) Does CoCo-Bot provide competitive generative quality while ensuring strict interpretability? (2) How effective is CoCo-Bot at enabling composable, explicit, and direct concept interventions?

4.1. Concept Accuracy and FID

Table 1 reports both the concept classification accuracy and Fréchet Inception Distance (FID) for our method compared to the previous CC-AE [8] baseline. Concept accuracy is measured as the average agreement between the pseudo-labels inferred from generated images and the intended concept composition. FID is used to assess the realism and diversity of generated samples.

CoCo-Bot achieves improved concept accuracy and competitive FID, indicating that it can faithfully realize user-specified concept interventions while maintaining high image quality. The gain in concept accuracy highlights the effectiveness of removing auxiliary vision cues—ensuring that all generative changes are directly attributable to explicit, human-understandable concepts.

4.2. Composability and Interventions

Figure 1 showcases the fine-grained editing capabilities of CoCo-Bot for a variety of concept-based interventions on CelebA-HQ [4] images. When a single semantic concept—such as "Mouth Open," "Smile," or "Makeup"—is activated, CoCo-Bot produces precise, visually consistent edits that are immediately interpretable and localized to the intended attribute. For instance, introducing "Mouth Open" to a source image reliably produces an open mouth without altering unrelated facial features.

Crucially, CoCo-Bot maintains strong compositionality even in more complex, multi-concept interventions. When users compose several concepts—for example, simultaneously activating "Smile," "Attractive," and negating "Male"—the model consistently generates images where each semantic attribute is faithfully and independently reflected. These interventions do not introduce unwanted entanglement or visual artifacts, which are often observed in previous CBM-based generative models that include auxiliary vision cues.

Counterfactual interventions, such as toggling a single attribute (*e.g.*, switching "Male" to "¬Male" while holding other concepts fixed), result in predictable and targeted changes, highlighting the model's controllability. Throughout all experiments, we observe that every edit is transparently and uniquely attributable to explicit concept intervention, and that there is no evidence of leakage from non-interpretable channels.

These qualitative findings underscore that CoCo-Bot achieves robust, transparent, and user-driven compositional editing. The method supports both simple and complex interventions, consistently translating user intent into visually coherent and semantically disentangled outputs.

5. Conclusions

We have presented CoCo-Bot, an energy-based and composable concept bottleneck framework for interpretable generative models. In contrast to prior approaches that rely on an auxiliary vision cue at the bottleneck, CoCo-Bot constrains the generative process to operate solely through explicit, human-understandable concepts. By leveraging energy-based modeling and diffusion-style score matching, our method enables robust, compositional, and fully interpretable concept interventions, without compromising generative quality.

Empirical results on CelebA-HQ demonstrate that CoCo-Bot achieves higher concept accuracy and competitive FID compared to previous CBM-based generative baselines. Qualitative experiments further show that users can reliably compose, negate, and intervene on semantic concepts in a transparent and predictable manner.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [25ZD1120, Development of ICT Convergence Technology for Daegu-GyeongBuk Regional Industry] and funded by REXEN and ETRI [25RD1100, Development of On-device AI-based Video Analysis Technology]

References

[1] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In *NeurIPS*, pages 21400–21413, 2022. 1, 2

[2] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020. 2

[3] Aya Abdelsalam Ismail, Julius Adebayo, Hector Corrada Bravo, Stephen Ra, and Kyunghyun Cho. Concept bottleneck generative models. In *ICLR*, 2023. 1, 2

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 2, 4

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2

[6] Sangwon Kim, Dasom Ahn, Byoung Chul Ko, In-su Jang, and Kwang-Ju Kim. Eq-cbm: A probabilistic concept bottleneck with energy-based models and quantized vectors. In *ACCV*, pages 3432–3448, 2024. 1, 2

[7] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, pages 5338–5348, 2020. 1, 2

[8] Akshay Kulkarni, Ge Yan, Chung-En Sun, Tuomas Oikarinen, and Tsui-Wei Weng. Interpretable generative models through post-hoc concept bottlenecks. In *CVPR*, pages 8162–8171, 2025. 1, 2, 4

[9] Hankook Lee, Jongheon Jeong, Sejun Park, and Jinwoo Shin. Guiding energy-based models via contrastive latent variables. In *ICLR*, 2023. 2

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3

[11] Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *JMLR*, 4:1235–1260, 2003. 2

[12] Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *ICLR*, 2024. 1, 2