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

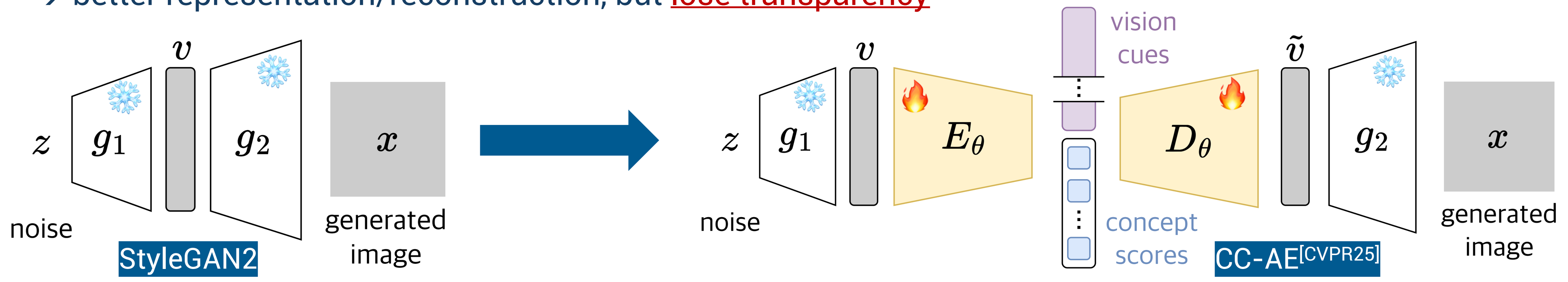
Sangwon Kim In-Su Jang Pyoung-Kun Kim Kwang-Ju Kim

ETRI – Electronics and Telecommunications Research Institute, South Korea

ICCV HONOLULU HAWAII OCT 19-23, 2025

1. Motivation

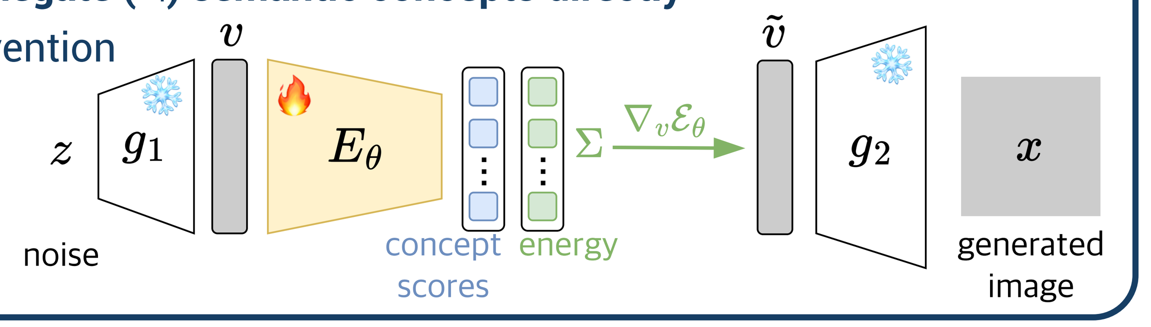
- Generative models need interpretability for trustworthy AI
- Prior Generative Concept Bottleneck Models (CBMs) use auxiliary vision cues and a decoder(D_θ)
  → better representation/reconstruction, but lose transparency



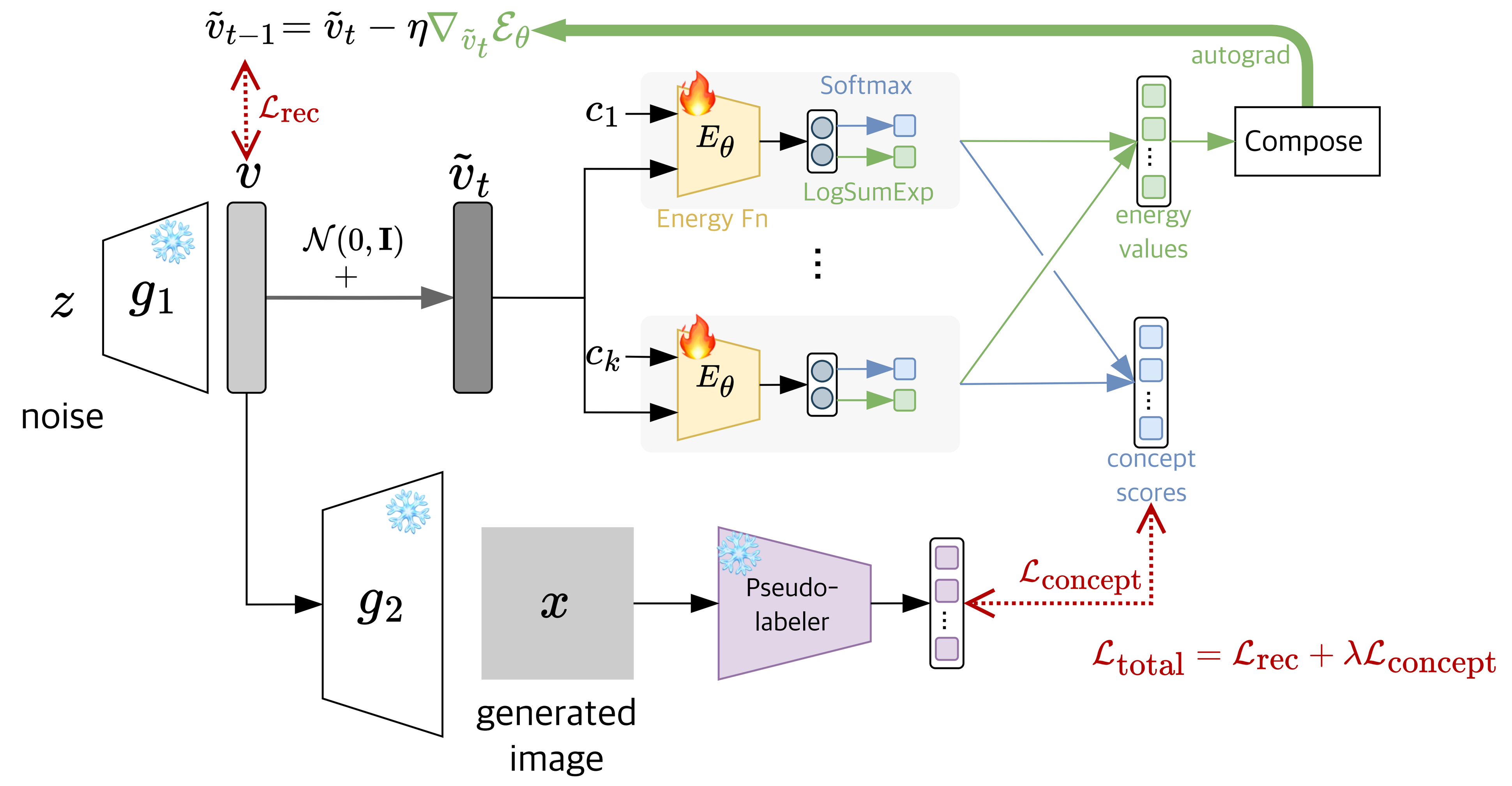
Can we achieve transparent, composable, and human-controllable generation without any auxiliary vision cues or a decoder?

Our Idea: CoCo-Bot

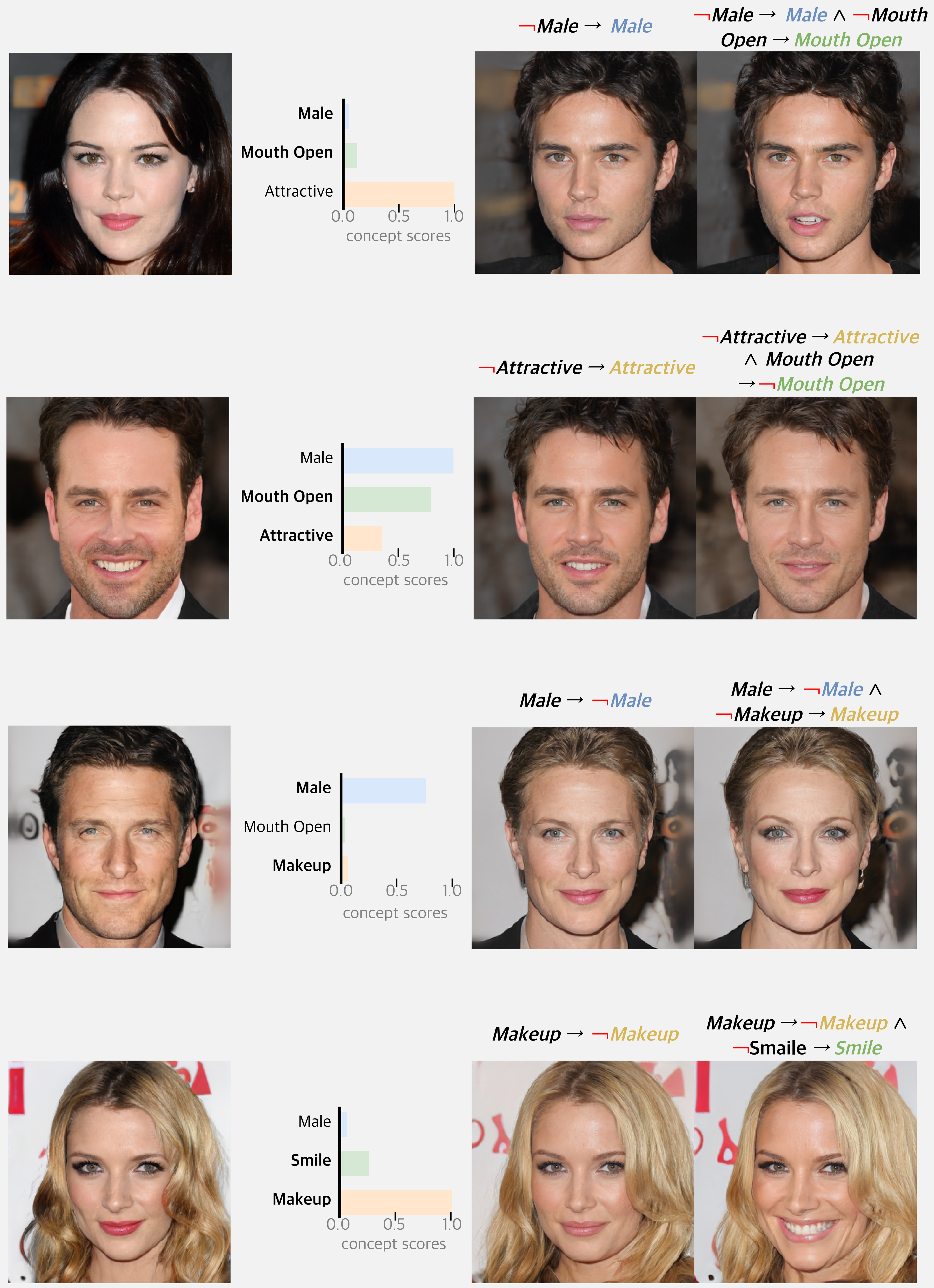
1) **Energy-guided Concept Bottleneck (Decoder-free)**
   → Overcoming information bottleneck, better transparency
2) **Diffusion-scheduled Energy Functions**
   → Better stable sampling than conventional MCMC/SGLD of energy-based models
3) **Composable Interventions: compose (∧) or negate (¬) semantic concepts directly**
   → Precise and disentangled semantic intervention



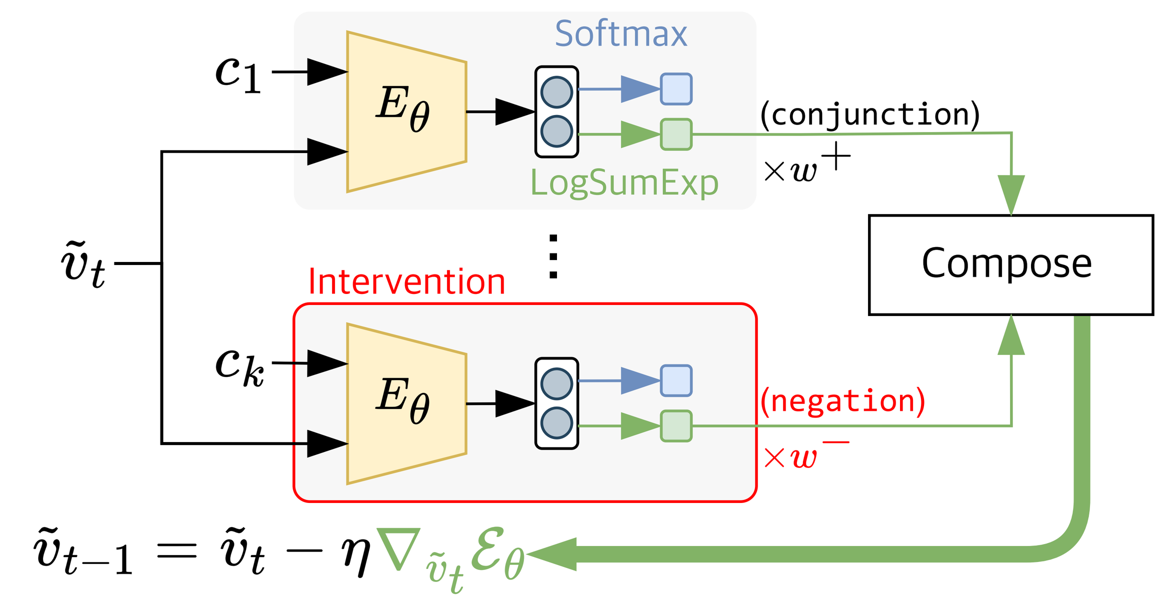
2. Training CoCo-Bot



Original *Concepts* *Interventions*



3. Composable Human-intervention



Users can directly compose or negate semantic concepts for transparent intervention

Human-intervention is applied through weights:
$w^+ (= 1.0)$ for conjunction
$w^- (= -0.001)$ for negation
enabling transparent concept composition

4. Quantitative Result

	CelebA-HQ		CUB	
Method	Concept Acc. (%, ↑)	FID (↓)	Concept Acc. (%, ↑)	FID (↓)
CC-AE[CVPR25]	74.38	9.77	75.56	8.37
CoCo-Bot (Ours)	**75.70**	**6.47**	**82.42**	**5.37**

Ours achieves **+6.86% Accuracy** & **-3.0 FID** (CUB) *vs.* CC-AE → Better interpretability & realism

5. Qualitative Result

Concept Interpretation
→ Single-concept Intervention
→ Multi-concept intervention
: explicit, controllable edits

CUB reconstruction:
Ours sharper & faithful
vs.
CC-AE artifacts

Original CC-AE Ours