PAPER REVIEW PRESENTATION

Mixture of Concept Bottleneck Experts (M-CBE)

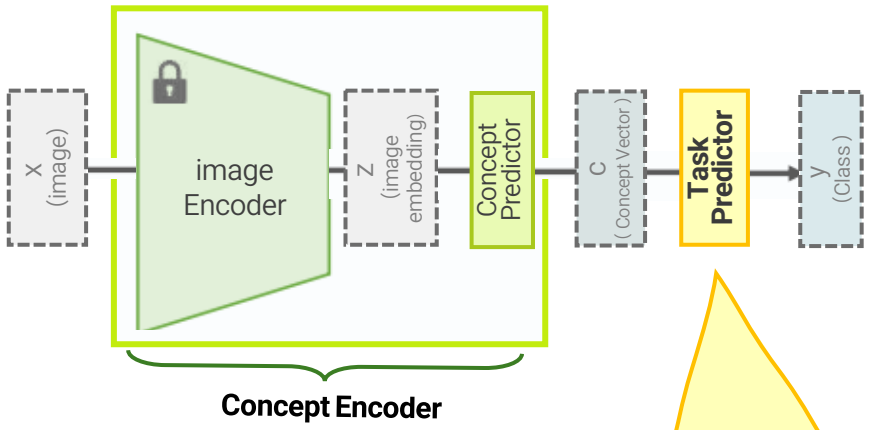
Francesco De Santis, Gabriele Ciravegna, Giovanni De Felice, and Arianna Casanova
ICML 2026 Spotlight

2026.7.16

SNU IDEA Lab. 정유립

01 Mixture of Concept Bottleneck Experts

CBM (2020)



limitation A
함수 형태가 하나로 미리 정해짐
ex:

$$[f(c) = w^{\top}c + b]$$

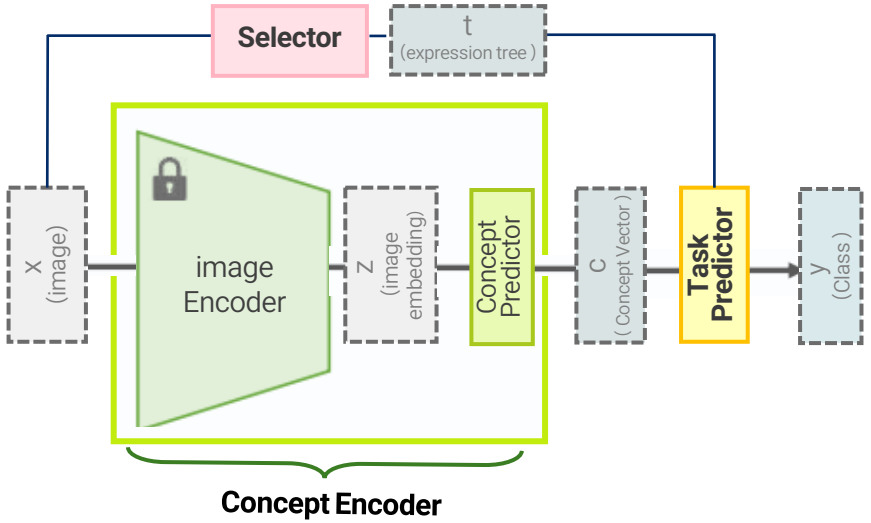
또는

$$[f(c) = \mathrm{MLP}(c)]$$

limitation B
모든 입력에 같은 함수 하나를 사용함

$$\left[x_1, x_2, \ldots \quad \rightarrow \quad \text{모두 같은 } f\right]$$

M-CBE



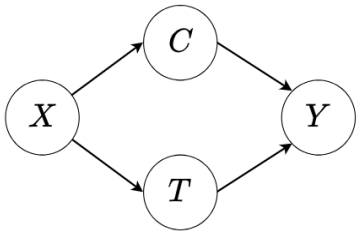
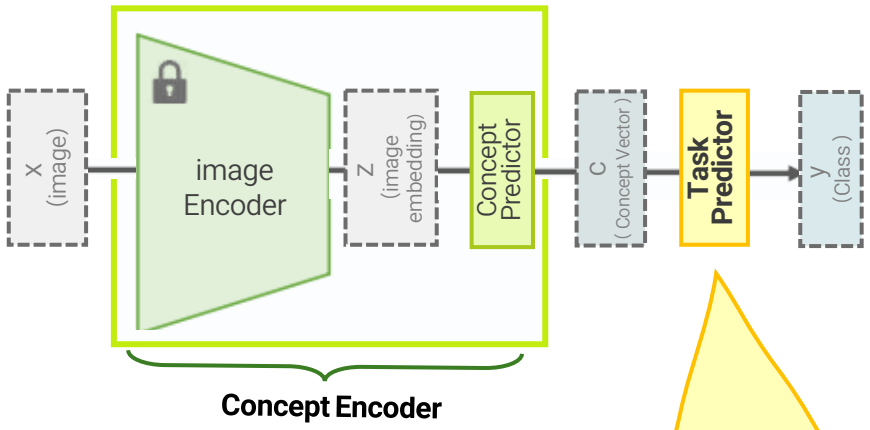


Figure 2. Left:
PGM representing the assumed generative processfor M-CBEs.

01 Mixture of Concept Bottleneck Experts

CBM (2020)



limitation A
함수 형태가 하나로 미리 정해짐
ex:

$$[f(c) = w^{\top}c + b]$$

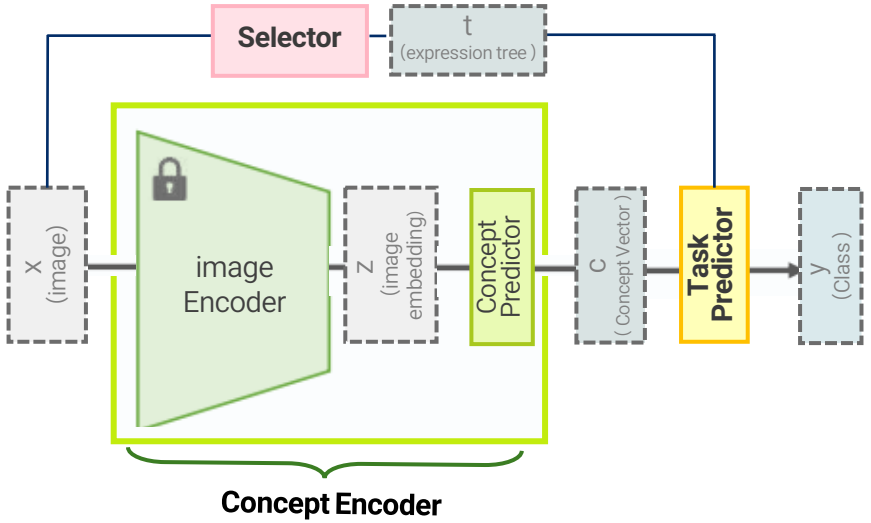
또는

$$[f(c) = \mathrm{MLP}(c)]$$

limitation B
모든 입력에 같은 함수 하나를 사용함

$$\left[x_1, x_2, \ldots \quad \rightarrow \quad \text{모두 같은 } f\right]$$

M-CBE



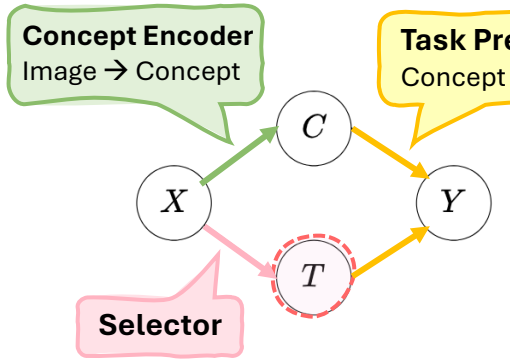


Figure 2. Left:
PGM representing the assumed generative processfor M-CBEs.

4. M-CBE 전체 구조
이제 처음으로 전체 구조를 보여줘.

01 Mixture of Concept Bottleneck Experts

Expression Tree

Example) $y = \theta_1 c_1 + \theta_2 c_2$

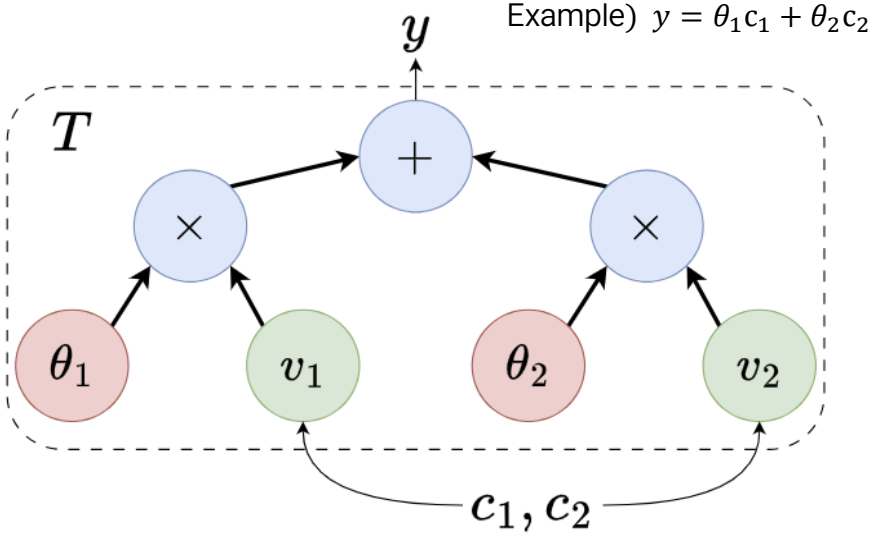


Figure 2. Right:
Example of an expression tree (T) representing a linear expression. Green nodes are placeholder variables V ,blue nodes are operators O drawn from the vocabulary W, andred nodes are learnable parameters Θ.
At inference time, each placeholder $v_i$is instantiated with the corresponding predictedconcept c_i

Selector & Symbolic regression

1. Hyperparameter, Tree (Expert) 개수 M 설정
2. M 만큼의 임시 MLP Expert $g_1, \ldots, g_M$ 를 Placeholder로 설정
3. Selector와 임시 MLP Expert를 end to end로 학습
4. 이후, 각 MLP를 잘 근사하는 Tree를 탐색 : PySR
5. MLP Expert $g_1, \ldots, g_M$ 를 탐색된 Tree $t_1, \ldots, t_M$으로 교체
6. 탐색된 Tree로 계수 fine tuning

01 Mixture of Concept Bottleneck Experts

M-CBE

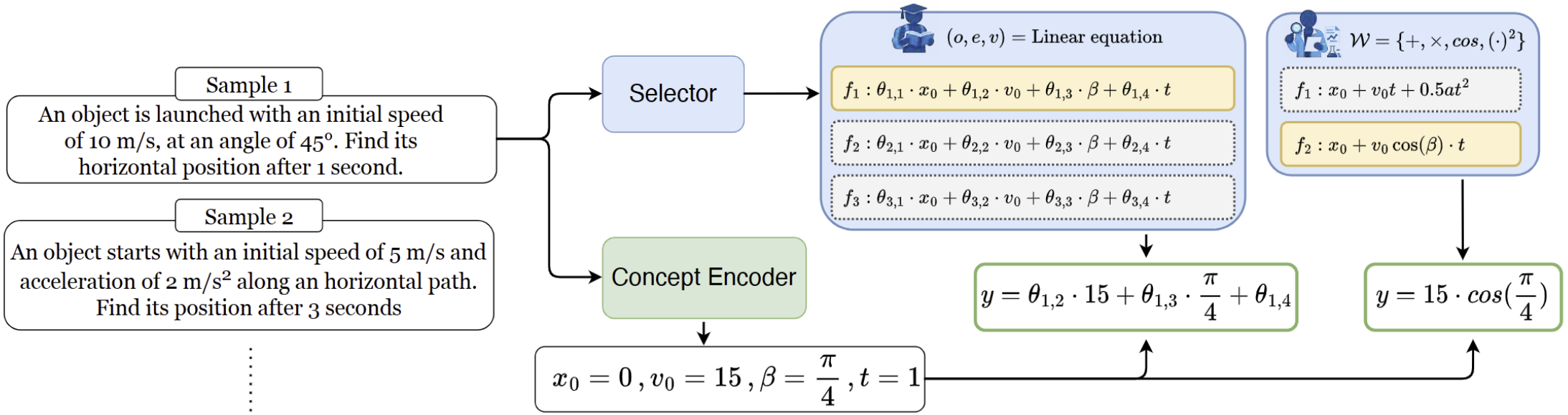


Figure 3. Overview of M-CBEs. The selector identifies the appropriate expression based on the input. Simultaneously, the Concept Encoder predicts the underlying physical concepts (x_0, v_0, α, t) from the raw input. The architecture accommodates different user mental models by allowing functional forms to be represented as either parametric linear equations or symbolic expressions restricted to interpretable operators. The final output y is obtained by executing the selected function with the predicted concepts as arguments.

02 Experimental Results

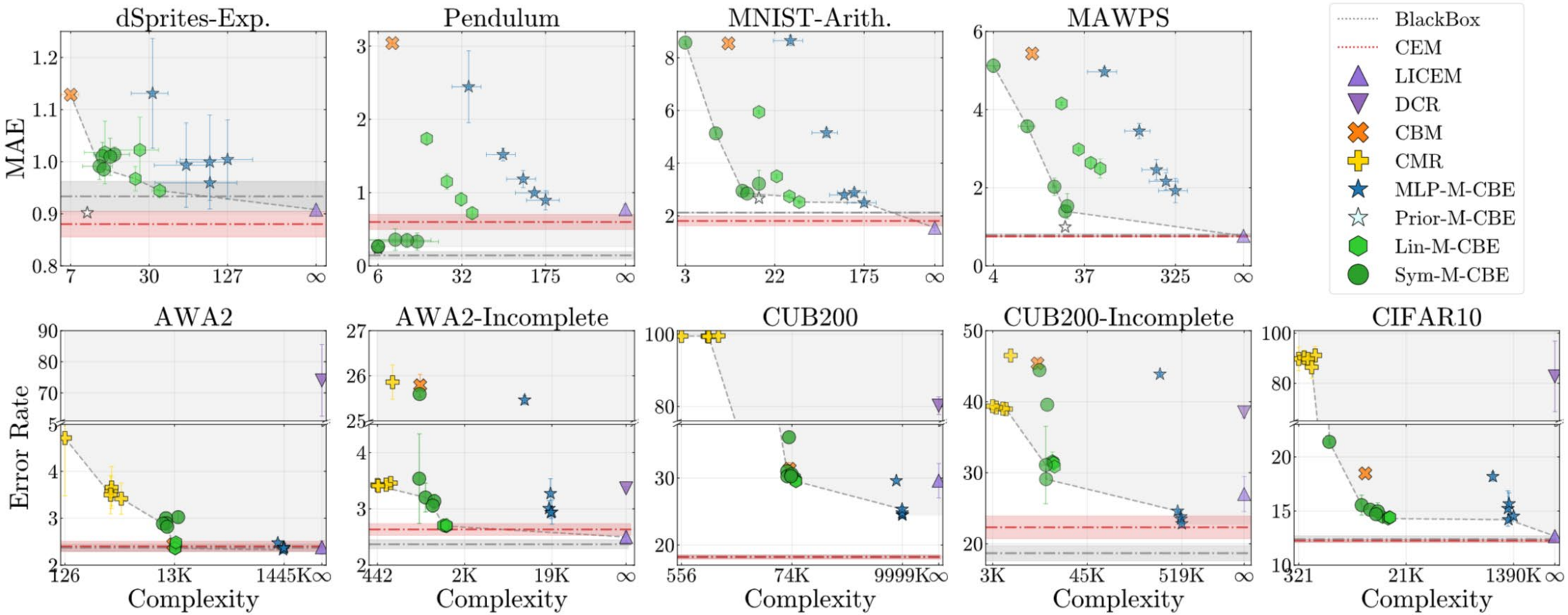


Figure 4. The x-axis denotes model complexity, measured as the total number of nodes across all expression trees used by the taskpredictor. The y-axis reports MAE for regression tasks (top) and error rate for classification tasks (bottom). For multi-expert models,multiple points are shown corresponding to different numbers of experts (1 – 5). The dotted curve indicates the Pareto frontier, while theshaded region marks dominated solutions. Prior-M-CBE is excluded from the Pareto frontier as it relies on ground-truth expressions.CEM and BlackBox are plotted as horizontal lines, since their label-predictor complexity is undefined (they do not operate on conceptpredictions). Error bars indicate 95% confidence intervals over five random seeds

02 Experimental Results

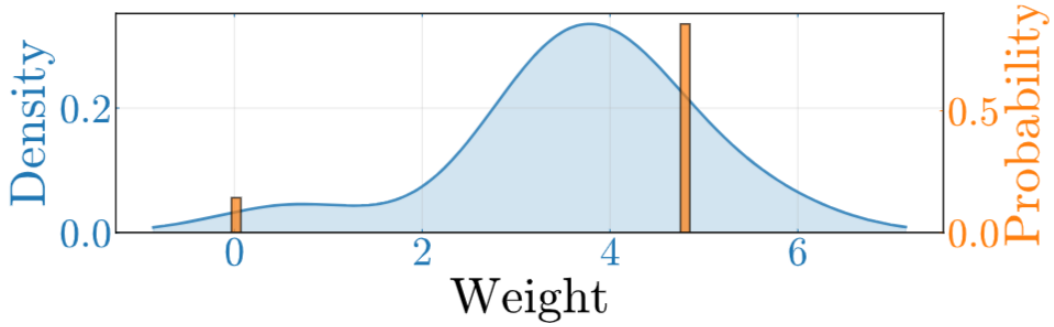


Figure 5. Weight distributions for the concept “has underpartscolor brown” toward class “Crested Auklet”, comparing LICEM(blue) and Lin-M-CBE (orange, memory size = 2).

Model	dSprites-Exp.	Pendulum	MNIST-Arith.	MAWPS
Lin-M-CBE	$18.00_{\pm 0.00}$	$5.71_{\pm 0.00}$	$4.71_{\pm 3.23}$	$15.00_{\pm 0.00}$
MLP-M-CBE	$28.00_{\pm 0.00}$	$43.63_{\pm 0.00}$	$22.24_{\pm 1.75}$	$47.00_{\pm 0.00}$
Sym-M-CBE	$13.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.05_{\pm 0.09}$	$0.00_{\pm 0.00}$

Table 1. Tree Edit Distance (TED) to the ground truth expres-sion. In three datasets, Symbolic M-CBE recovers the exact data-generating mechanism (TED ≈ 0)

	Pendulum		MAWPS	
Model	MAE	Complexity	MAE	Complexity
MLP-M-CBE	$0.46_{\pm 0.08}$	$706.0_{<0.01}$	$1.05_{\pm 0.04}$	$3592.0_{<0.01}$
Sym-M-CBE (S)	$3.80_{\pm 0.02}$	$3.0_{<0.01}$	$2.73_{\pm 0.03}$	$4.0_{<0.01}$
Sym-M-CBE (M)	$2.12_{\pm 1.53}$	$19.3_{\pm 5.8}$	$1.29_{\pm 0.08}$	$24.0_{<0.01}$
Sym-M-CBE (C)	$0.46_{\pm 0.13}$	$6.0_{<0.01}$	$1.36_{\pm 0.21}$	$24.0_{<0.01}$

Table 2. Performance of Sym-M-CBE across small (S), medium(M), and complete (C) operator sets, measured by MAE and modelcomplexity (±95% confidence interval).

02 Experimental Results

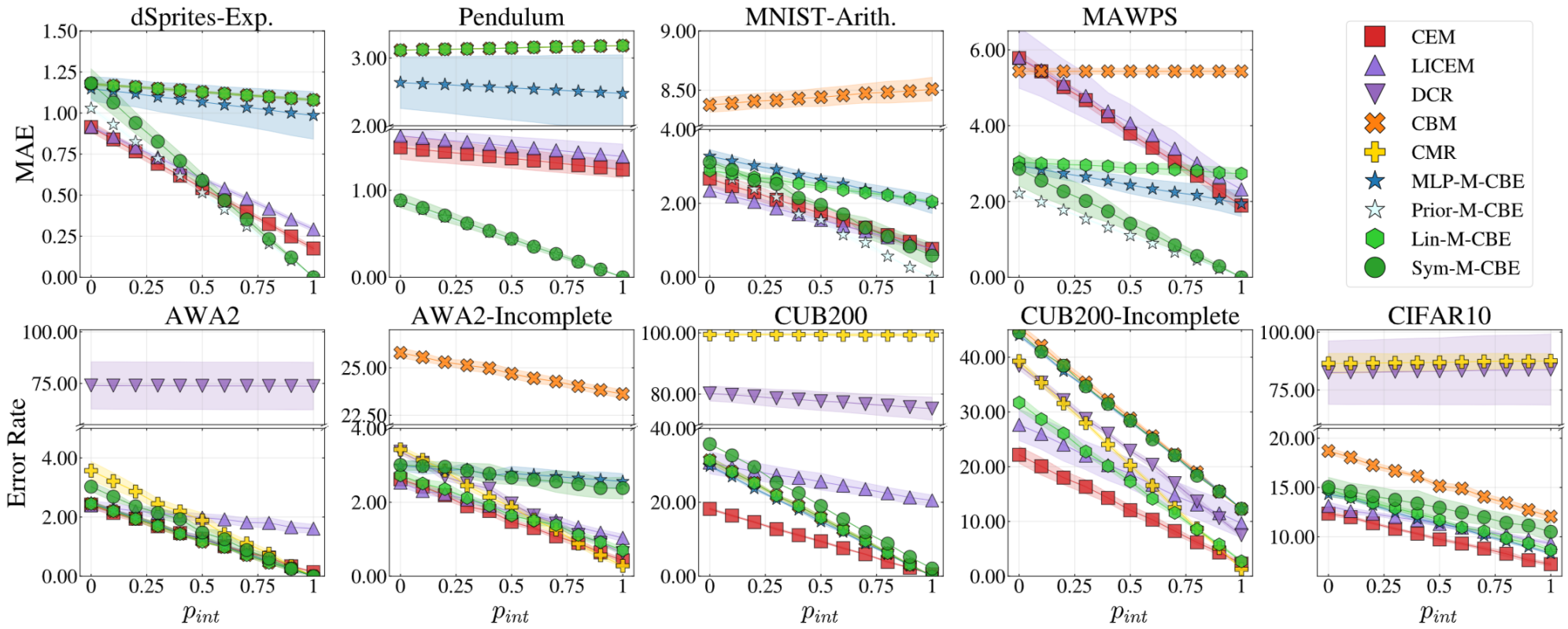
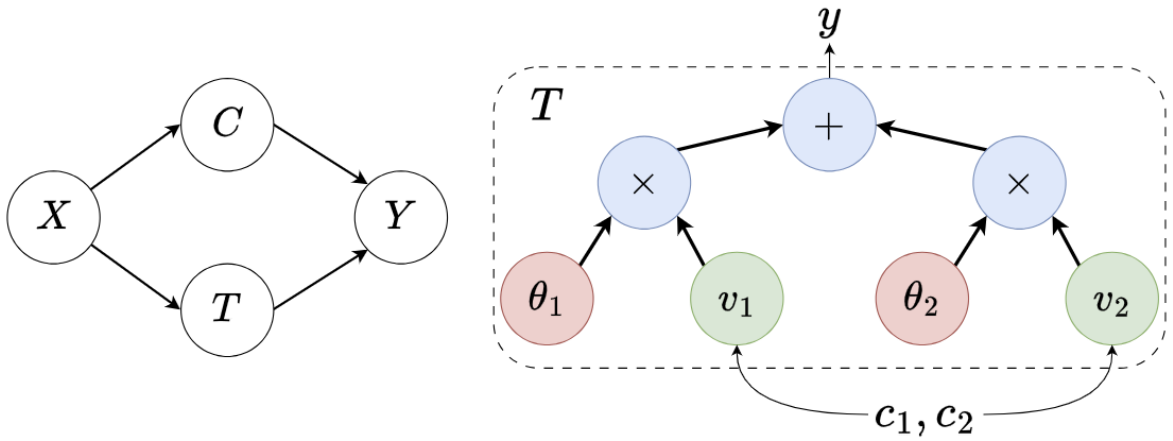


Figure 6. Effect of interventions on model performance. MAE (top) and error rate (bottom) as a function of intervention probability pint.Shaded areas show 95% confidence intervals over 5 seeds. For models with multiple experts, we use the configuration at the Pareto knee.The knee point is identified using the maximum distance to chord method：the point with maximum perpendicular distance to the lineconnecting the best complexity and best accuracy extremes, representing the optimal trade-off.



ure 2, right). We define a class $\mathcal{T}$ of expression trees w.r.t. a vocabulary of operations $\mathcal{W}$ via admissible edge configurations $\mathcal{E}$ on a set of nodes $\mathcal{N} = \mathcal{V} \cup \mathcal{O} \cup \mathcal{P}$, with:
(i) $\mathcal{V}$: set of *input nodes* instantiated with specific **concepts**.
(ii) $\mathcal{O}$: set of *operator nodes* (e.g., $\{\times, \sin, \exp\}$) from $\mathcal{W}$.
(iii)$\mathcal{P}$: set of *parameter nodes* representing real numbers.

Figure 2. **Left:** PGM representing the assumed generative process for M-CBEs. **Right:** Example of an expression tree (T) representing a linear expression. Green nodes are placeholder variables V, blue nodes are operators O drawn from the vocabulary $\mathcal{W}$, and red nodes are learnable parameters Θ. At inference time, each placeholder v_i is instantiated with the corresponding predicted concept c_i.