南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Concept Bottleneck Model

Pang Wei Koh [*1] Thao Nguyen [*1 2] Yew Siang Tang [*1]
Stephen Mussmann [1] Emma Pierson [1] Been Kim [2] Percy Liang [1]

ICML 2020

Background

End-to-end: X → Y

CBM: X → C → Y

任务准确率与概念可解释性之间是否存在权衡? 即，模型在完成具体任务（如分类、预测）时的准确度，是否会因为增加对中间概念的解释能力而降低，或者是否可以兼顾这两者

测试时的干预（interventions）是否有助于提升模型的准确度? 也就是说，在模型运行时通过人为修改模型预测的中间概念，是否能够改进模型最终的预测结果。

概念准确率（concept accuracy）是否是有效进行干预能力的良好指标? 这里探讨的是，如果模型能准确预测中间概念，那么对这些概念做出的干预是否更可能带来有效的改善。

不同的训练方法会不会导致干预效果有显著差异? 换句话说，不同的训练策略对概念瓶颈模型的干预效果有怎样的影响，这些差异是否重要。

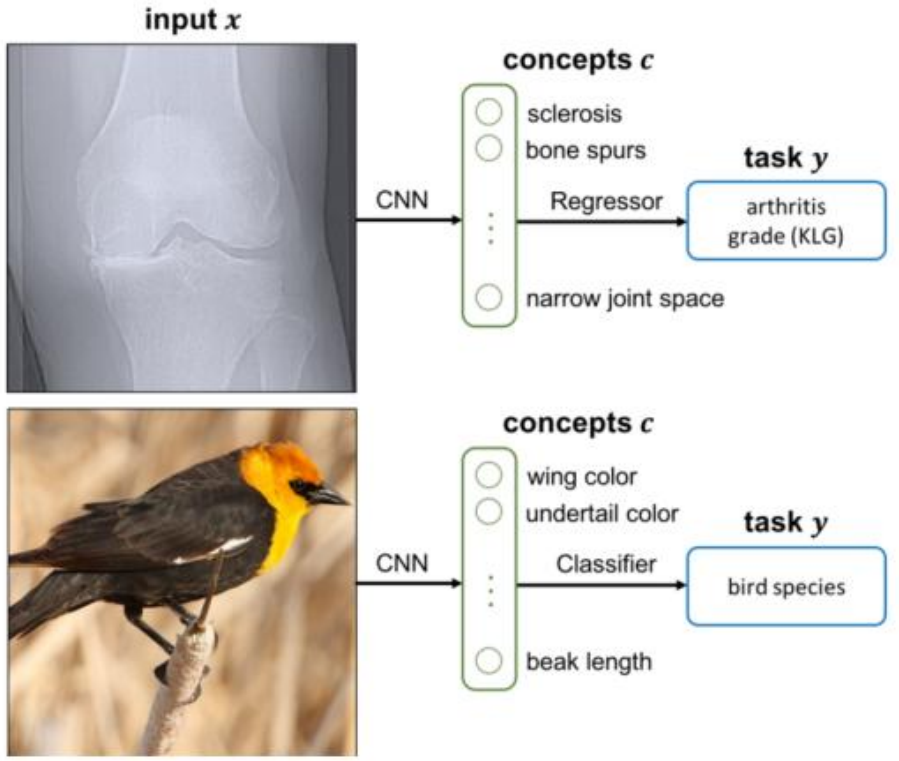


Figure 1. We study concept bottleneck models that first predict an intermediate set of human-specified concepts c, then use c to predict the final output y. We illustrate the two applications we consider: knee x-ray grading and bird identification.

Methods

$$X \rightarrow C \rightarrow Y$$

$\hat{c} = g(x)$，$\hat{y} = f(\hat{c})$　　$\hat{y} = f(g(x))$

- g : X → C，负责把原始输入映射到 k 维概念空间。
- f : C → Y，只使用预测概念完成最终任务预测。

- 将普通网络某一层调整为 k 个神经元，并用概念损失逐维对齐真实概念 c_j。
- C 不一定必须因果导致 Y；只要 C 与 Y 有稳定预测关系，CBM 仍可利用。
- 训练阶段需要概念标注；测试阶段只输入 x，由模型自动预测 ĉ。

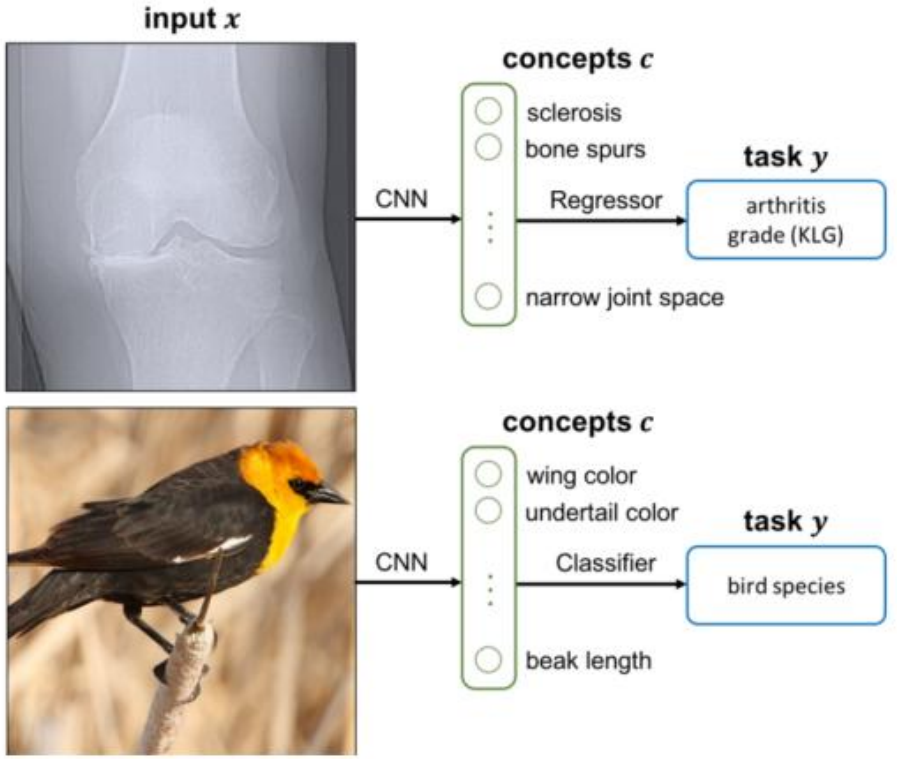


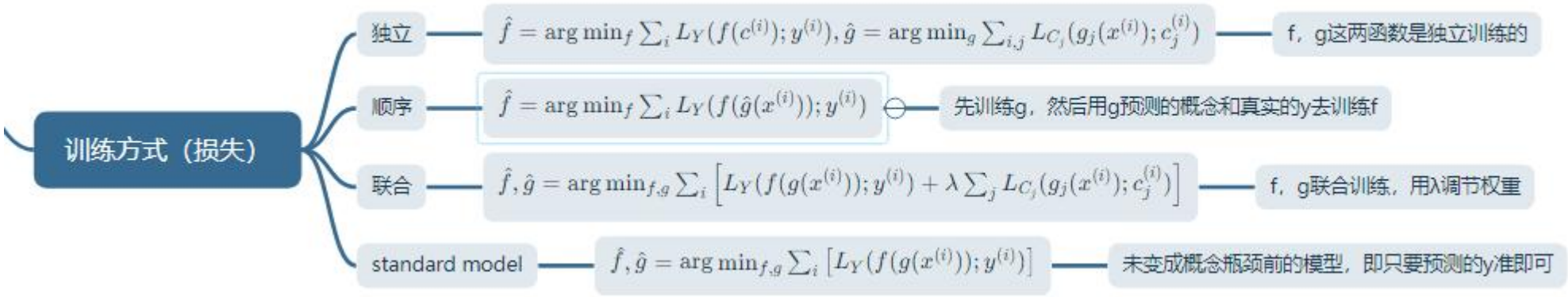
Figure 1. We study concept bottleneck models that first predict an intermediate set of human-specified concepts c, then use c to predict the final output y. We illustrate the two applications we consider: knee x-ray grading and bird identification.

Methods

$$X \rightarrow C \rightarrow Y$$

$\hat{c} = g(x)$，$\hat{y} = f(\hat{c})$　　$\hat{y} = f(g(x))$

- g : X → C，负责把原始输入映射到 k 维概念空间。
- f : C → Y，只使用预测概念完成最终任务预测。



Methods

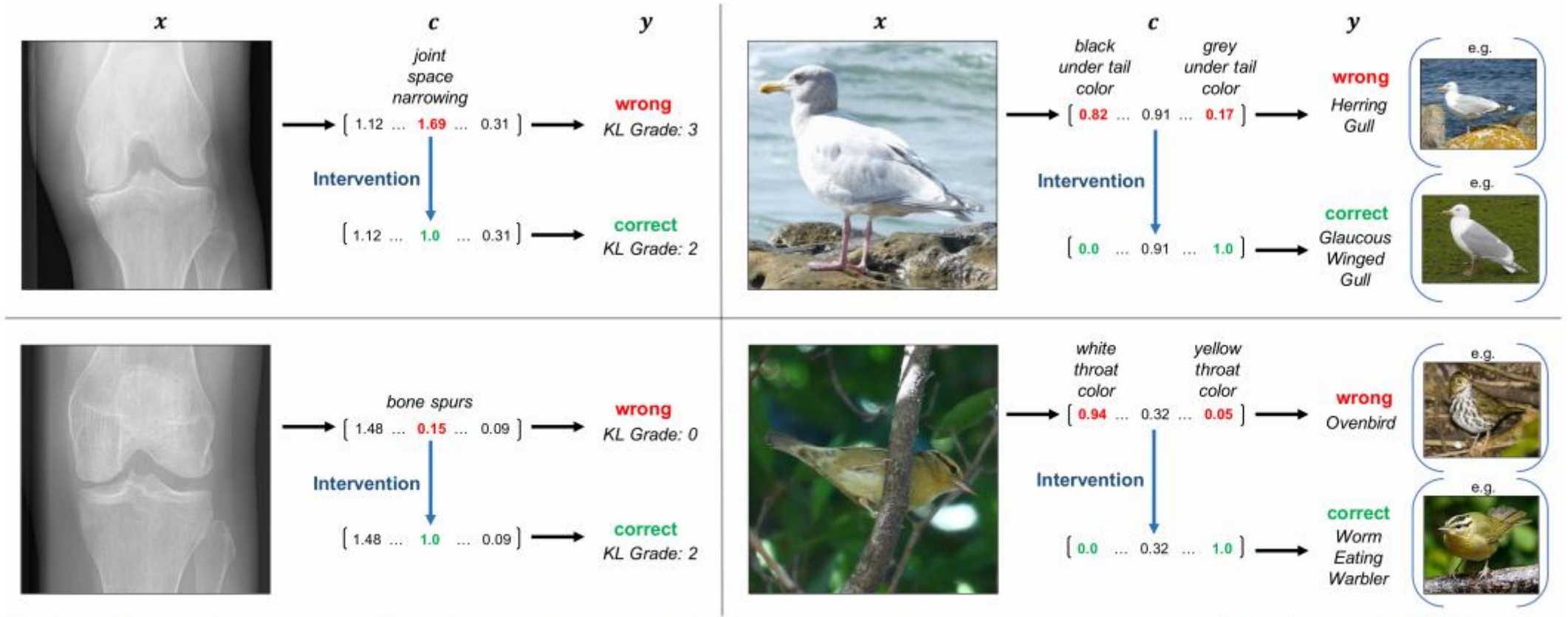


Figure 3. Successful examples of test-time intervention, where intervening on a single concept corrects the model prediction. Here, we show examples from independent bottleneck models. **Right**: For CUB, we intervene on concept groups instead of individual binary concepts. The sample birds on the right illustrate how the intervened concept distinguishes between the original and new predictions.

Experiments

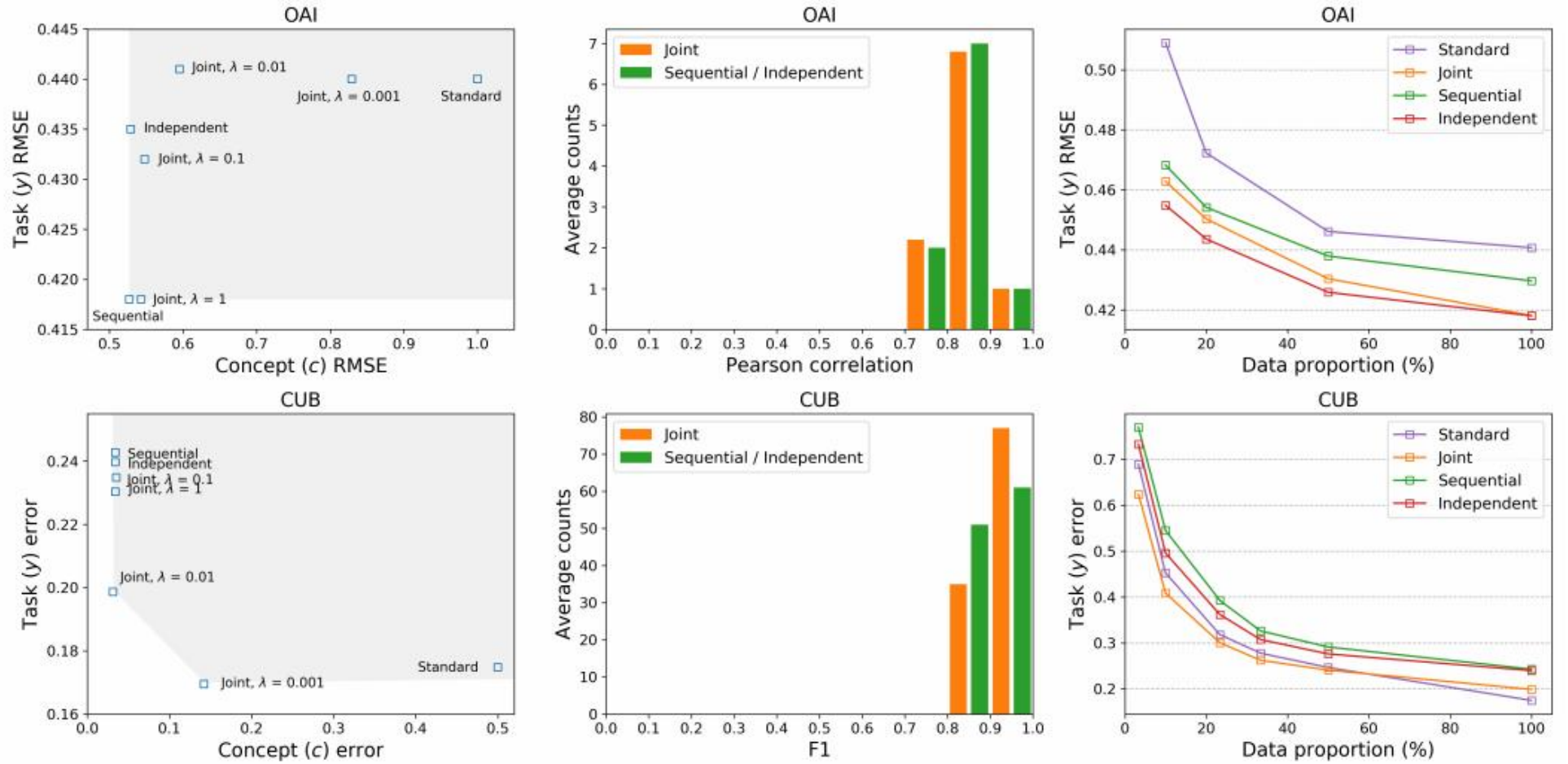


Figure 2. **Left**: The shaded regions show the optimal frontier between task vs. concept error. On OAI, we find little trade-off; models can do well on both task and concept prediction. On CUB, there is some trade-off, with standard models and joint models that prioritize task prediction (i.e., with sufficiently low λ) having lower task error. For standard models, we plot the concept error of the mean predictor (OAI) or random predictor (CUB). **Mid**: Histograms of how accurate individual concepts are, averaged over multiple random seeds. In our tasks, each individual concept can be accurately predicted by bottleneck models. **Right**: Data efficiency curves. Especially on OAI, bottleneck models can achieve the same task accuracy as standard models with many fewer training points.

Experiments

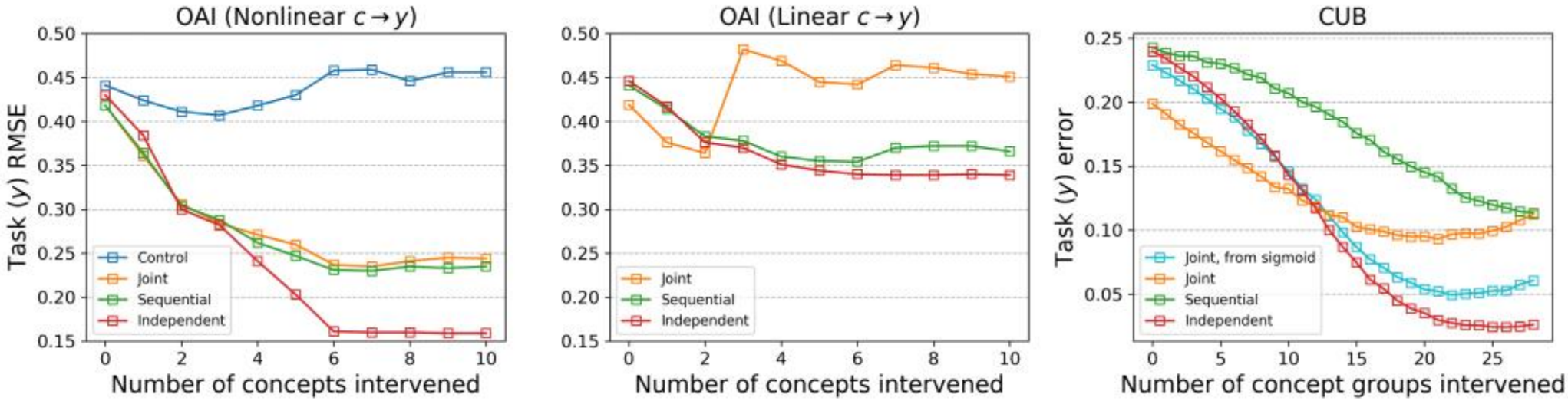


Figure 4. Test-time intervention results. **Left**: Intervention substantially improves task accuracy, except for the control model, which is a joint model that heavily prioritizes label accuracy over concept accuracy. **Mid**: Replacing $c \to y$ with a linear model degrades effectiveness. **Right**: Intervention improves task accuracy except for the joint model. Connecting $c \to y$ to probabilities rescues intervention but degrades normal accuracy.

Discussion

- 概念标注成本：CBM 训练数据需要 (x, c, y)，比普通监督学习多出 concept annotation。
- 概念质量问题：concept 标错、噪声大、概念集合不完整，都会影响最终预测和干预效果。

Thanks