How Well Do Vision Models Encode Diagram Attributes?

Haruto Yoshida [1,*], Keito Kudo [1,2], Yoichi Aoki [1,2], Ryota Tanaka [1,3], Itsumi Saito [1], Keisuke Sakaguchi [1,2], Kentaro Inui [4,1,2]

[1] Tohoku University, [2] RIKEN, [3] NTT Human Informatics Laboratories, [4] MBZUAI [*] yoshida.haruto.p1@dc.tohoku.ac.jp

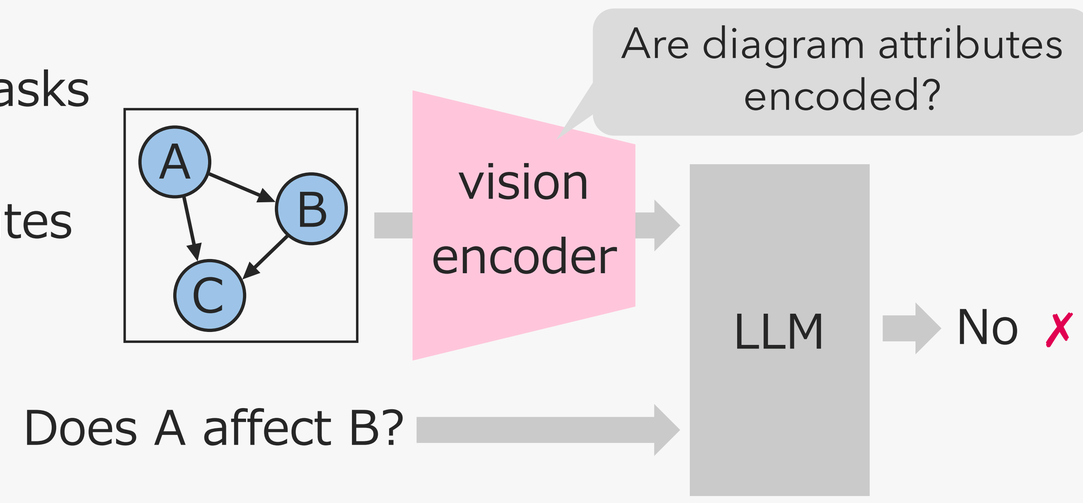
Summary

- We probed **how vision models encode diagram attributes** such as node shape and edge direction.
- Vision models **do not encode attributes like edge direction into a low-dimensional subspace**.

Background

- Vision encoders are employed for some diagram-related tasks (e.g., diagram understanding).
- It's unclear whether vision models encode diagram attributes (e.g., node color, edge direction).

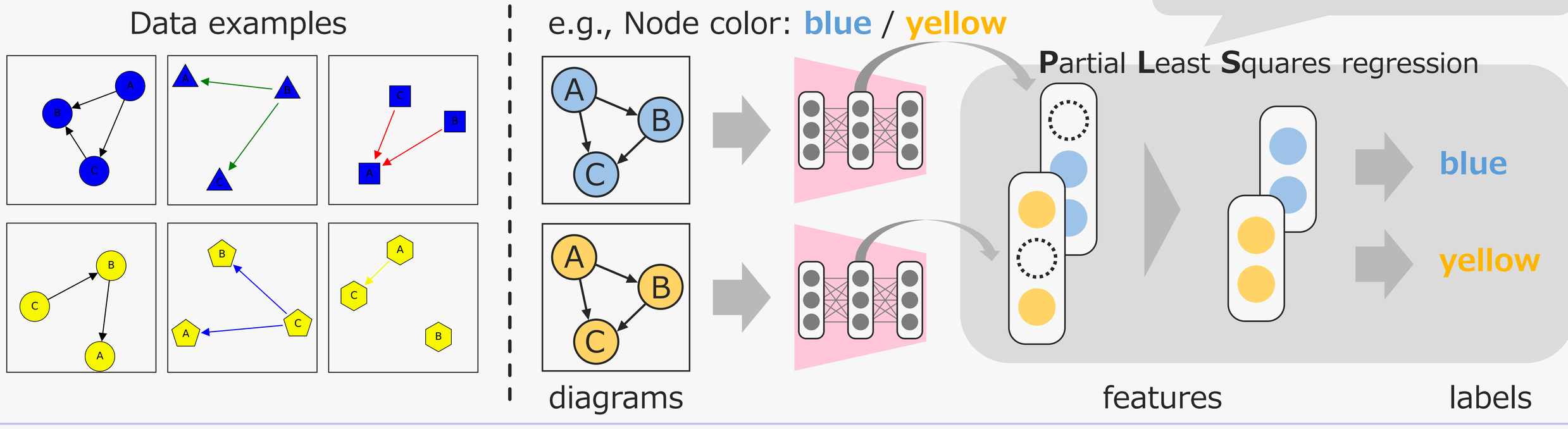
Investigating **if internal representations retain diagram attributes** (=Probing)



Dataset & Probing procedure

- We constructed a dataset consisting of **directed graph-based diagrams**.
- We performed **binary classification for each attribute** (e.g., node color) and evaluated its accuracy.

Intentionally extracting key information for classification



Experiment

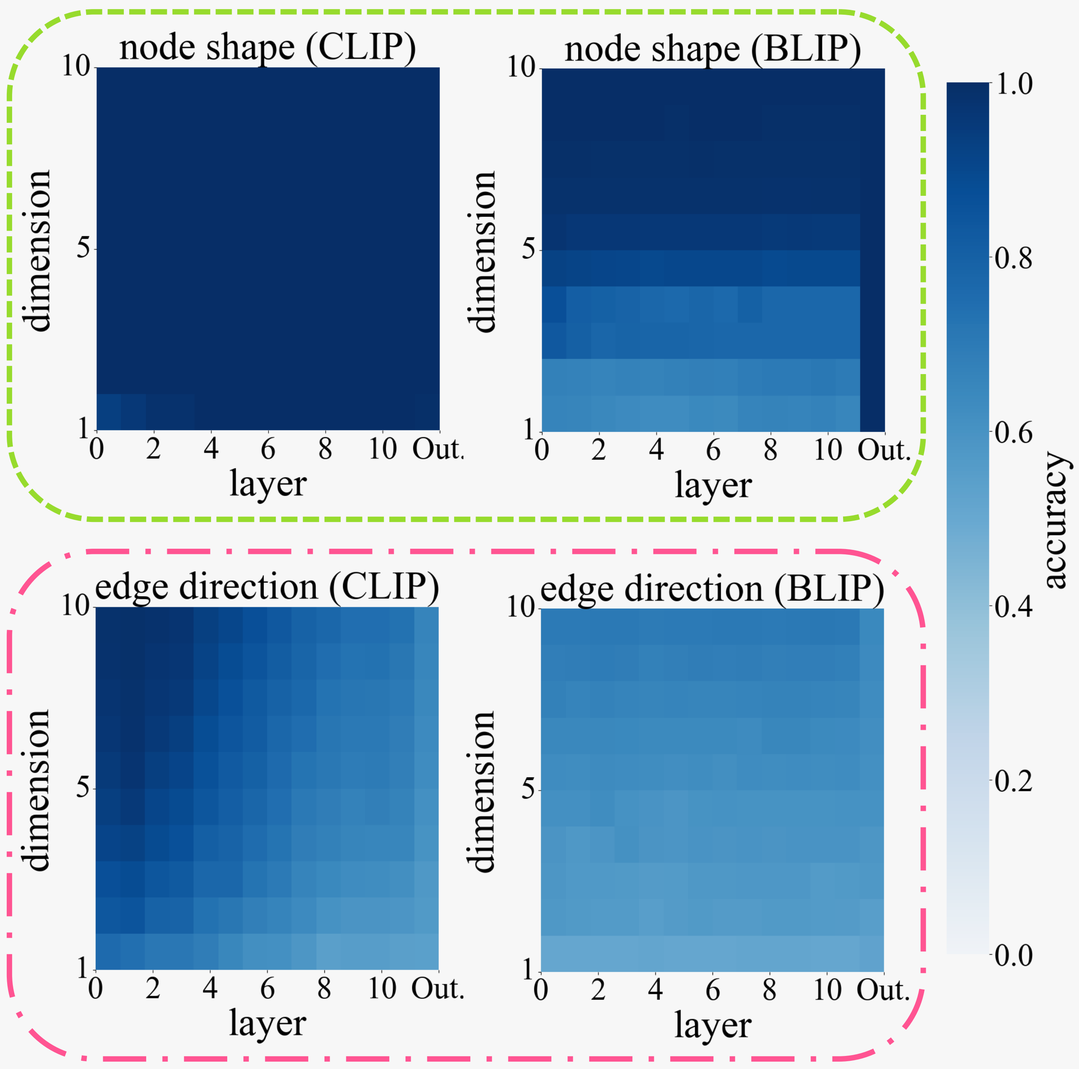
- Models: CLIP [Radford+ICML'21], BLIP [Li+ICML'22]
- Target attributes:

○ node color
○ node shape
○ edge color

Encoded in a low-dimensional linear subspace

○ edge existence
○ edge direction

Not encoded in a low-dimensional linear subspace



Why is it difficult to encode edge direction?

- Pre-training data lacks **high-quality diagram-text pairs**.
  - Only a few diagrams are included.
  - Even fewer diagrams come with detailed descriptive text.

We plan to **train vision models with diagrams**.