(ICML 2026)
How can embedding models bind concepts?

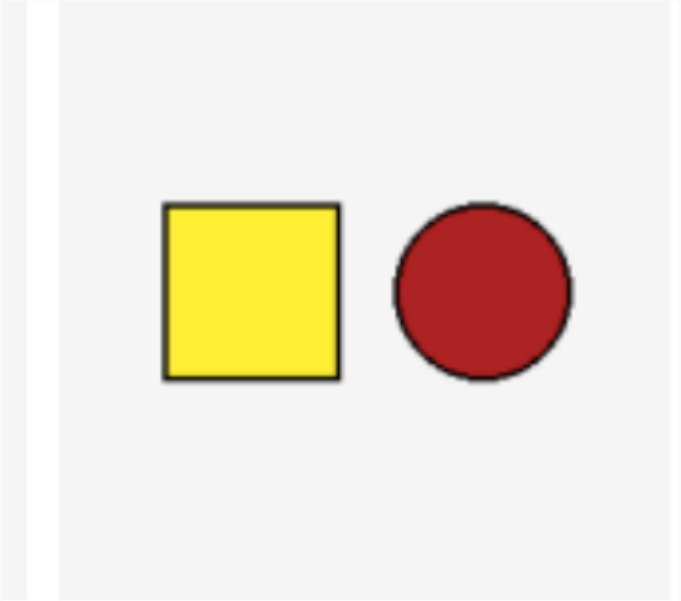
Uselis et al.

Presented by Hankyo Jeong

July 16, 2026

Seoul National University

Binding problem



A. a **yellow square** and a **red circle** ✓

B. a **red square** and a **yellow circle** ×

Concept binding: identifying which concepts belong to the same object.

Question

Observed tension	
Cross-modal	Image–text matching often fails to distinguish the correct pairing.
Uni-modal	A probe on image or text embeddings recovers object information.

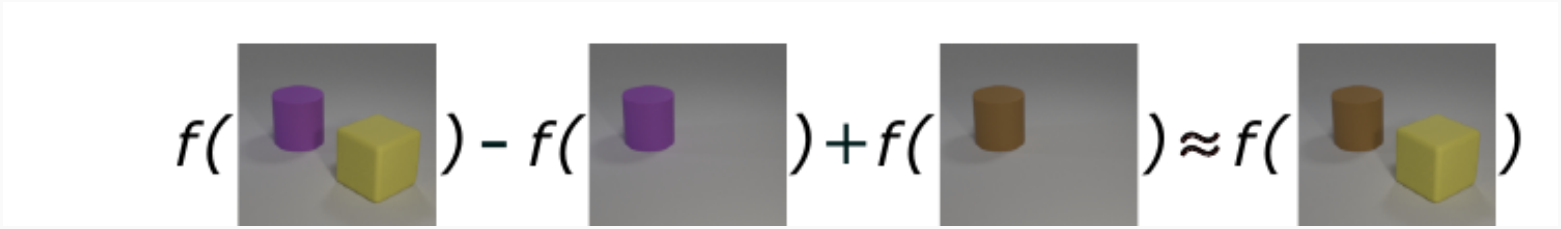
Object information is present,

but it is not aligned across modalities.

Analysis 1: scene embedding $\approx$ sum of object components

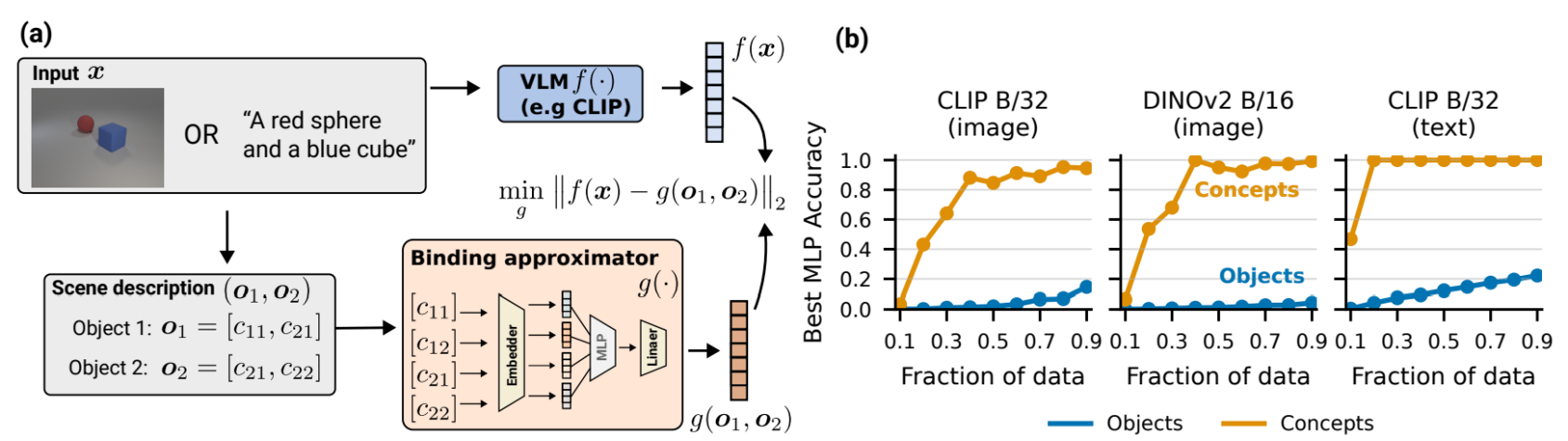
$$z_{\text{scene}} \approx u_{\text{object 1}} + u_{\text{object 2}}$$

$$z_{\text{image}} \approx u_{\text{yellow square}} + u_{\text{red circle}}$$



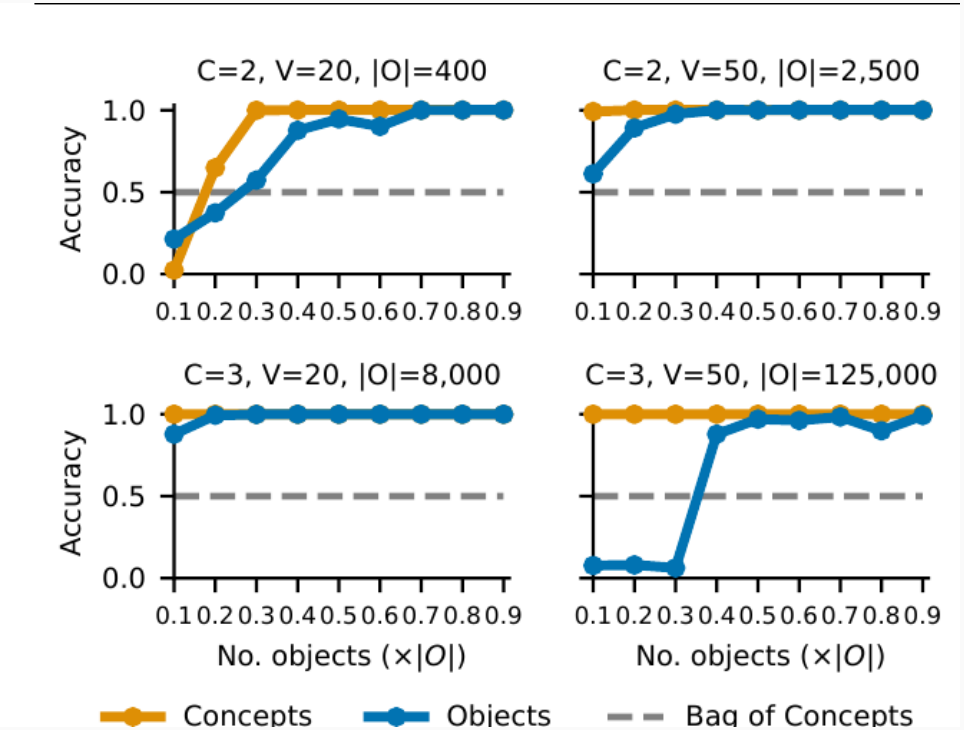
Object-level components are present in the embedding.

Analysis 2: does CLIP's binding rule generalize?



Concepts generalize; objects do not.

Is the failure fundamental?



Controlled experiment

Train CLIP-style dual encoders from scratch on controlled synthetic multi-object data.

- Test objects are **unseen concept combinations**.
- Concepts generalize early.
- With enough coverage, object recognition reaches high accuracy.

Embedding models can learn generalizable binding.

What structure explains generalizable binding?

Adding concepts is not enough

$$A : \text{yellow} + \text{square} + \text{red} + \text{circle},$$

$$B : \text{red} + \text{square} + \text{yellow} + \text{circle}.$$

$$A = B$$

Addition captures which concepts are present, but not how they are paired.

Pairing-sensitive interaction

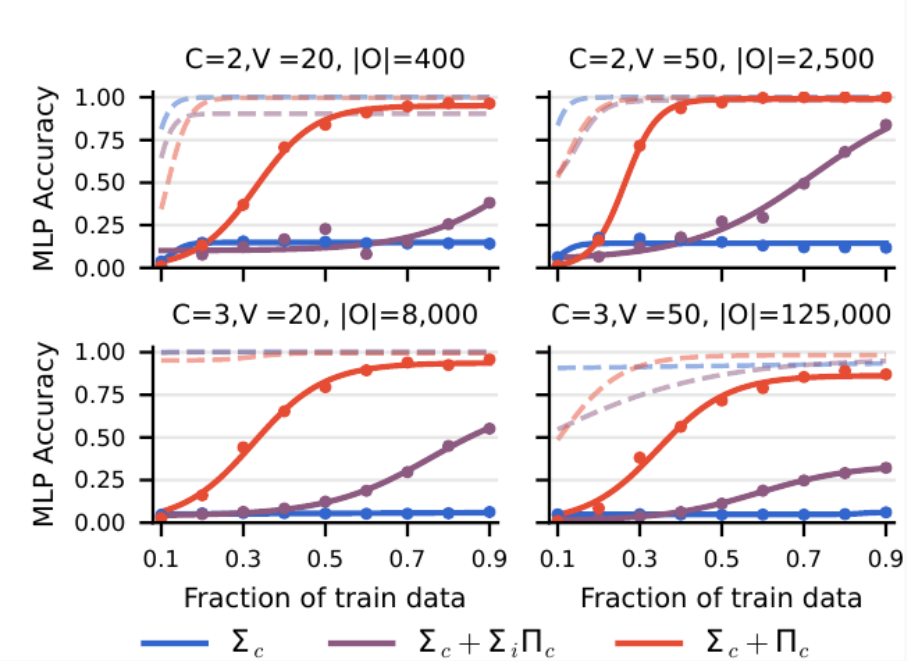
$$A : (\text{yellow} \times \text{square}) + (\text{red} \times \text{circle}),$$

$$B : (\text{red} \times \text{square}) + (\text{yellow} \times \text{circle}).$$

$$A \neq B$$

Multiplicative terms can distinguish object-level combinations.

Multiplicative structure best explains binding



- Additive probe: concepts ✓, objects ×.
- Additive + product probes: object recognition improves on unseen combinations.

Generalizing models are well described by a low-complexity multiplicative form.

Thank you!