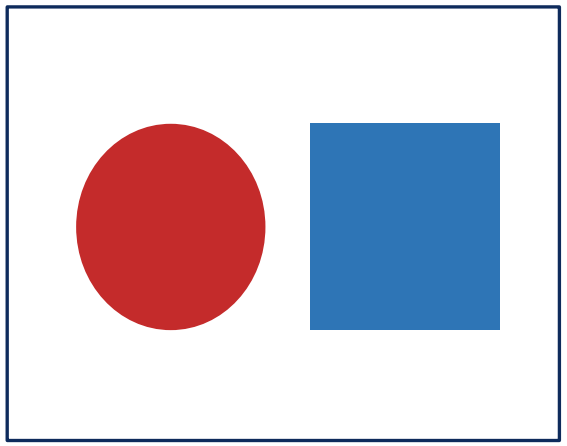
(ICML 2026)
How can embedding models bind concepts?

Author: Uselis et al.
Presented by Hankyo Jeong
2026. 07.22

Cross-Modal Binding Failure

Image x



- color = {red, green, blue}
- shape = {circle, triangle, square, ⋯, ⋯}

} concept

- object = (color, shape) concept pair
- o_1 = (red, circle), o_2 = (blue, square)
- scene = (o_1, o_2)

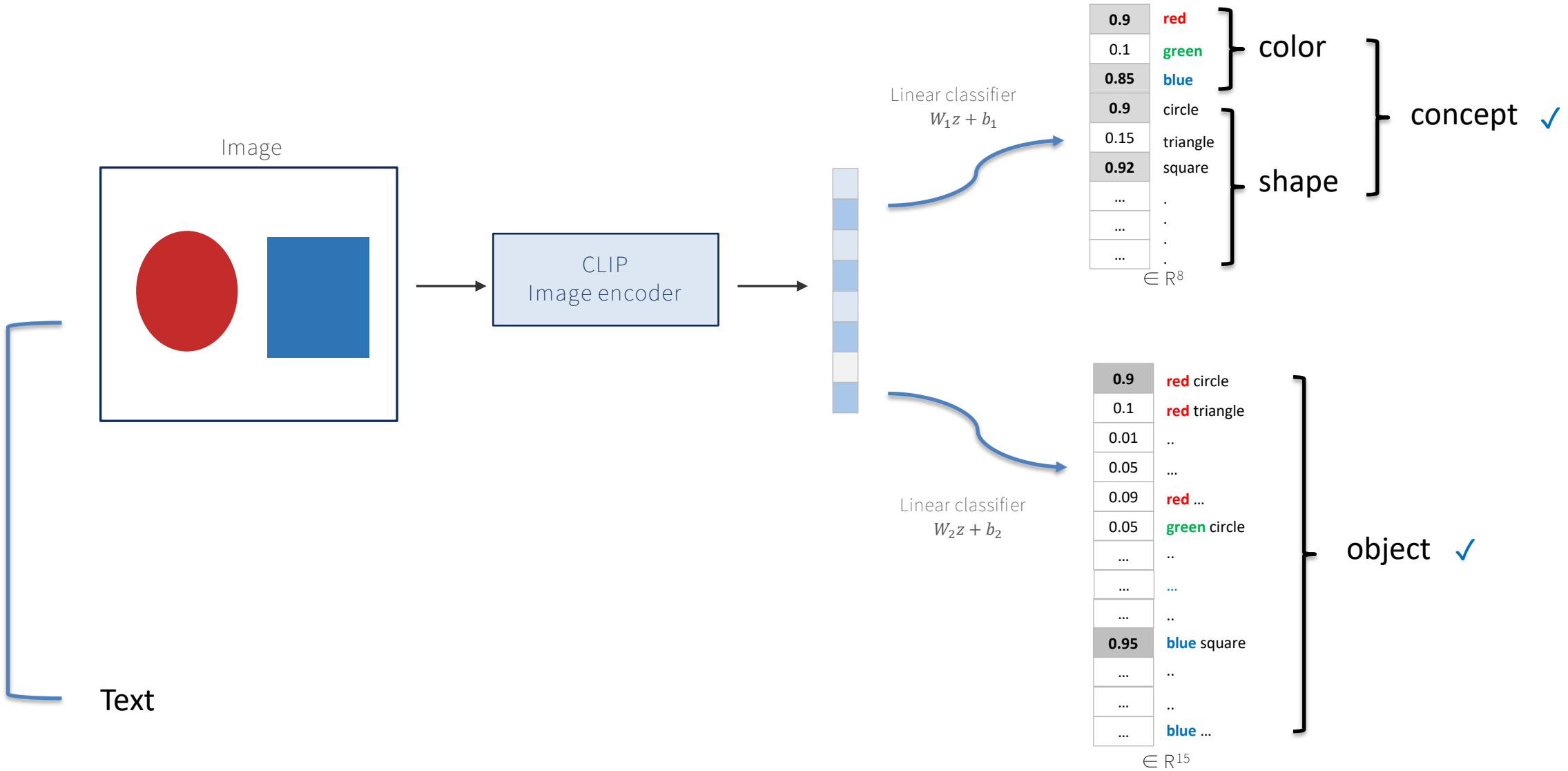
Caption A (correct ✔)
"A red circle and a blue square."

Caption B (incorrect ✘)
"A blue circle and a red square."

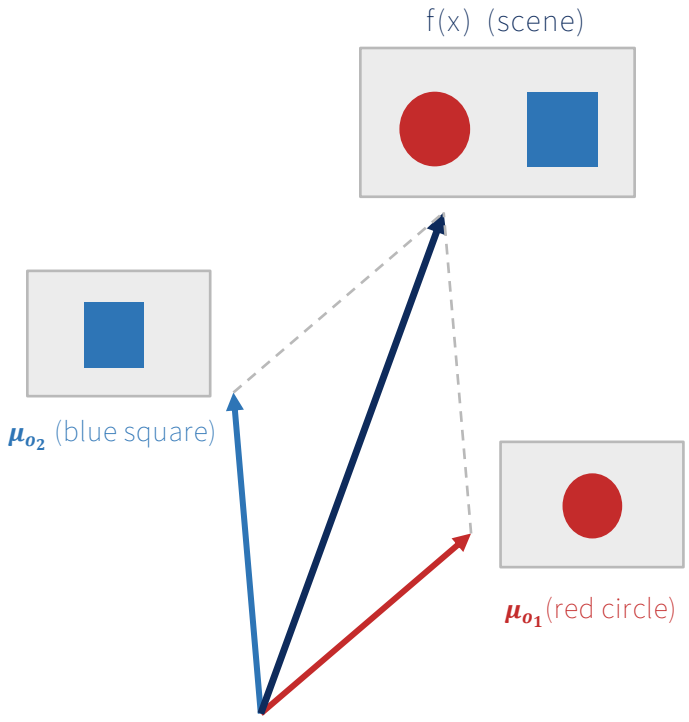
$\text{sim}(x, A) \approx \text{sim}(x, B) \longrightarrow$ Cross-modal binding failure

Uni-Modal Probing

Q. **Do CLIP embeddings fail to encode object information?**



Analysis 1. Geometric Structure

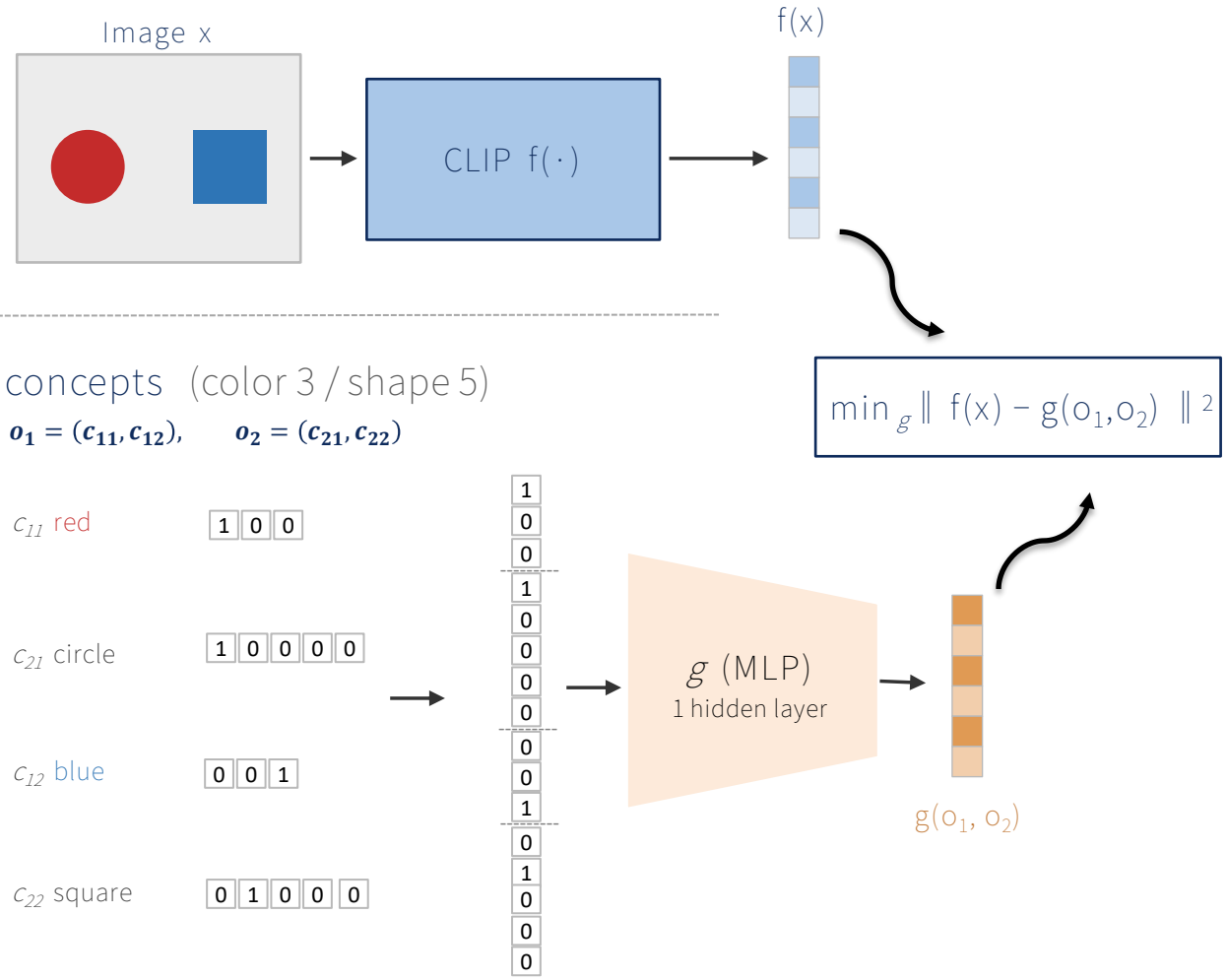


$$f(x) \approx \mu_{o_1} + \mu_{o_2}$$

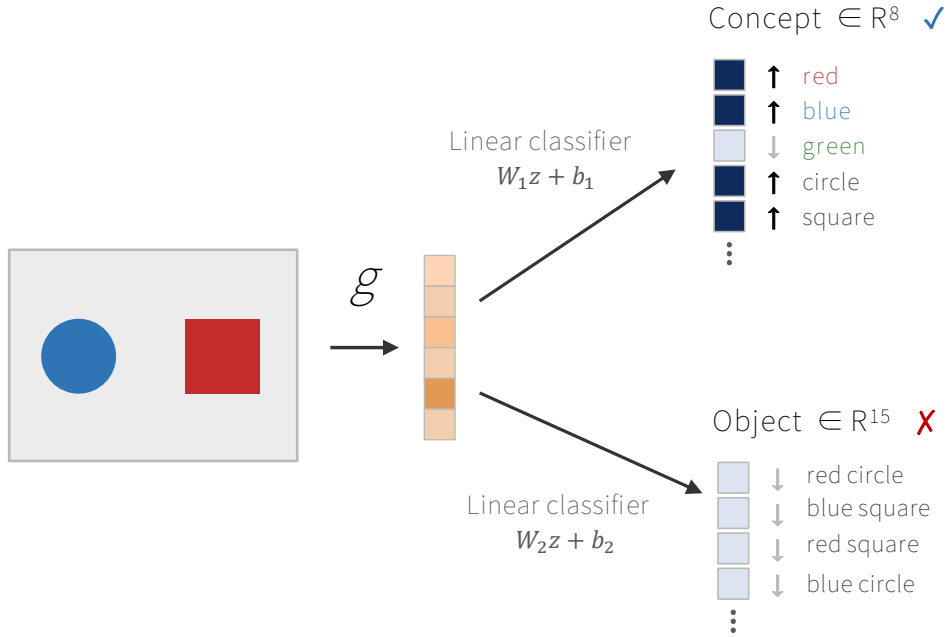
- For each modality (consider image)
  1. embed all images.
  2. Estimate object prototypes by averaging single-object embeddings.
  3. Visualize the embedding geometry using MDS.
  4. Confirm the additive structure in the original embedding space.

Analysis 2. Binding rule complexity

Train



Test



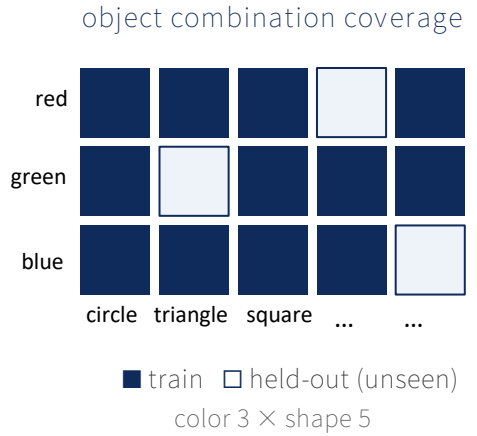
CLIP's binding rule is high-complex

Controlled Experiment — Architecture Limit or Data Problem?

Q. In a dual-encoder like CLIP, is binding generalization to unseen objects impossible?

Controlled setup

- Build a synthetic dataset covering all concept combinations
- Re-train both encoders from scratch
- Repeat the analysis → check cross-modal alignment



- Result: with enough coverage, binding generalizes to unseen combinations & cross-modal alignment succeeds ✓

Conclusion Not a limit of the dual-encoder architecture.
Existing CLIP failed to learn the binding rule due to insufficient object-combination coverage in training data

Future Work

Idea.

- 특정 domain dataset에서 concepts를 추출하고,
- 가능한 concept combinations으로 synthetic image–caption data를 생성하여 CLIP을 fine-tuning하면,
- unseen combinations에 대한 cross-modal binding 능력을 향상시킬 수 있을까?

Thank you!