NeurIPS 2026

NEURAL INFORMATION PROCESSING SYSTEMS

Can the Parts Fool the Test?

Counterfactual Pairing Cycles for Relational OOD



Yibo Zhou
Beihang University



Yirong Xiang
Air Force Medical Center

Counterfactual pairing cycles

Two images, two captions, four scores

	A person rides a horse	A person rides a bicycle
	**Matched** $+ s_{1\ 1}$	Swapped $- s_{1\ 2}$
	Swapped $- s_{2\ 1}$	**Matched** $+ s_{2\ 2}$

Only the pairing changes

Component shortcuts cancel exactly

The cycle contrast

$$\Delta = s_{1\ 1} + s_{2\ 2} - s_{1\ 2} - s_{2\ 1}$$

For any image-only and text-only functions

$$s(u, v) = a(u) + b(v)$$

$$\Delta = 0$$

$$\text{CycleAcc} = 0.5$$

Exact cancellation within every complete block

Pair-dependent artifacts can still remain

One contrast for learning and diagnosis

Cycle training

$$\mathscr{L} = \text{softplus}(-\Delta)$$

Additive heads cannot improve this objective

Bilinear diagnostic

$$\Delta = (\varphi_1 - \varphi_2)^\top M (\psi_1 - \psi_2)$$

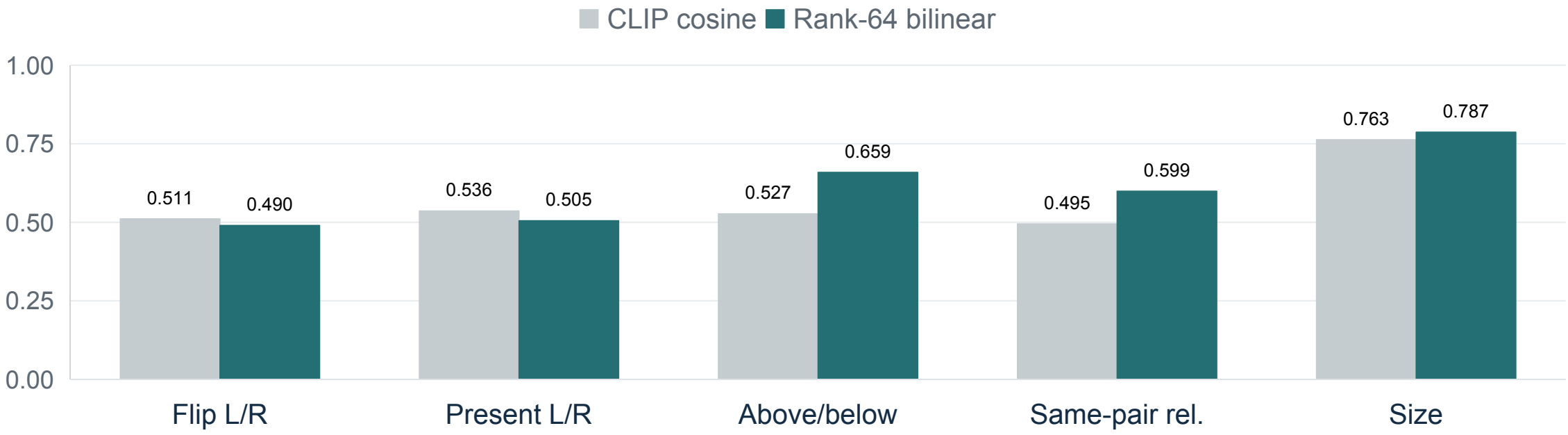
φ: image features ψ: text features

Rank tests which relation differences the features expose

Strict relations expose a binding gap

Both object categories are present in each target image

Exact CycleAcc



Left/right stays near chance. Size remains recoverable.

Frozen CLIP ViT-B/32. Means shown. Chance = 0.5. Full values ± SD in notes.

Shortcut validation overstates relational transfer

COCO outlier-exposure stress test

0.9998

Shortcut validation AUROC

A fixed prefix marks negative captions

Unmarked exact-cycle test

Training objective	CycleAcc
Shortcut OE	0.8817 ± 0.0021
Exact-cycle BCE	0.8853
Cycle OE	**0.9270 ± 0.0075**

+4.53 percentage points with Cycle OE

Compared with Shortcut OE on the same exact-cycle test. Mean ± SD across 3 seeds.

Relational OOD needs a test of the pairing

Keep the parts fixed.
Change the pairing.

Exact additive cancellation
A shared contrast for evaluation and training

Official implementation

ybzhou@buaa.edu.cn