NEURAL INFORMATION PROCESSING SYSTEMS

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad
॥ नवोन्मेषेण देशनिर्माणम् ॥

DiL
Data-driven Intelligence & Learning Lab

NeurIPS 2026

LACE: Latent Alignment via Counterfactual Embeddings

Harsh Udai* Konda Reddy Mopuri

Department of Artificial Intelligence, IIT Hyderabad, India

*Corresponding author

Pre-trained VLMs: Successes and Challenges

VLMs like CLIP have strong coarse-grained recognition and retrieval:

- Zero-shot classification and image–text retrieval.

However, they often fail at *attribute binding*, *role assignment* and *relational reasoning*.

✓ a red suitcase next to a black backpack

✗ a black suitcase next to a red backpack

Identical objects and attributes, different scene.

Web captions enumerate entities but leave bindings underspecified, so the loss admits co-occurrence shortcuts.

Preliminaries: CLIP-Style Contrastive Learning

Image and caption embeddings $\mathbf{v}_i$, $\mathbf{u}_i$ are ℓ_2-normalized, with scaled cosine similarity $s(\mathbf{v}, \mathbf{u}) = \mathbf{v}^\top \mathbf{u}/\tau$

Training minimizes the symmetric InfoNCE loss over a batch of *N* pairs:

$$\mathcal{L}_{\mathrm{I2T}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s(\mathbf{v}_i, \mathbf{u}_i))}{\sum_{j=1}^{N} \exp(s(\mathbf{v}_i, \mathbf{u}_j))}, \qquad \mathcal{L}_{\mathrm{T2I}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s(\mathbf{u}_i, \mathbf{v}_i))}{\sum_{j=1}^{N} \exp(s(\mathbf{u}_i, \mathbf{v}_j))}$$

Effective for global alignment, but it under-specifies the fine-grained structure needed for compositional reasoning.

Hard Negatives in CLIP

Hard negatives are *near-miss* distractors that sharpen contrastive boundaries.

Text-side (T-HN)

Perturbed captions $\tilde{u}_{i,k}$ (swaps, relation flips, negations) as extra negatives for image → text.

Cheap; used by most methods.

Image-side (I-HN)

Counterfactual images as extra negatives for text → image.

Expensive: editing, synthesis or mining.

Some methods instead use an auxiliary ranking loss, which pushes $\gamma\Delta_{i,k} \ll 0$:

$$\mathcal{L}_{\text{rank}} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \log\left(1 + \exp(\gamma\, \Delta_{i,k})\right), \qquad \Delta_{i,k} = s(\mathbf{v}_i, \tilde{\mathbf{u}}_{i,k}) - s(\mathbf{v}_i, \mathbf{u}_i)$$

Why Asymmetry Matters

Text-only hard negatives create an asymmetric training signal:

IMAGE → TEXT

Fine-grained pressure

✓ a red suitcase next to a black backpack

✗ a black suitcase next to a red backpack

✗ a red suitcase behind a black backpack

TEXT → IMAGE

No visual near-miss

✓ the matching image

only unrelated batch images as negatives

The image branch never has to separate near-identical scenes.

This can increase hubness, reduce reciprocity and destabilize neighborhoods, even when Recall@K improves.

Beyond Recall@K: Neighborhood Reciprocity

High recall can coexist with a warped topology. Is the shared space truly symmetric?

Mutual reciprocity (pair-level)

The true pair must be in each other's top-*K*, simultaneously.

$$\mathrm{Mutual@}K = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big[\big(i \in \mathcal{N}_K^I(i)\big) \wedge \big(i \in \mathcal{N}_K^T(i)\big)\Big]$$

Edge reciprocity (graph-level)

Share of all *K*-NN links that are bidirectional.

$$\mathrm{EdgeRecip@}K = \frac{\big|\{(i,j) \in E_{\mathrm{I2T}} : (j,i) \in E_{\mathrm{T2I}}\}\big|}{NK}$$

A large gap between Recall and Mutual@*K* means the space is poorly calibrated across modalities.

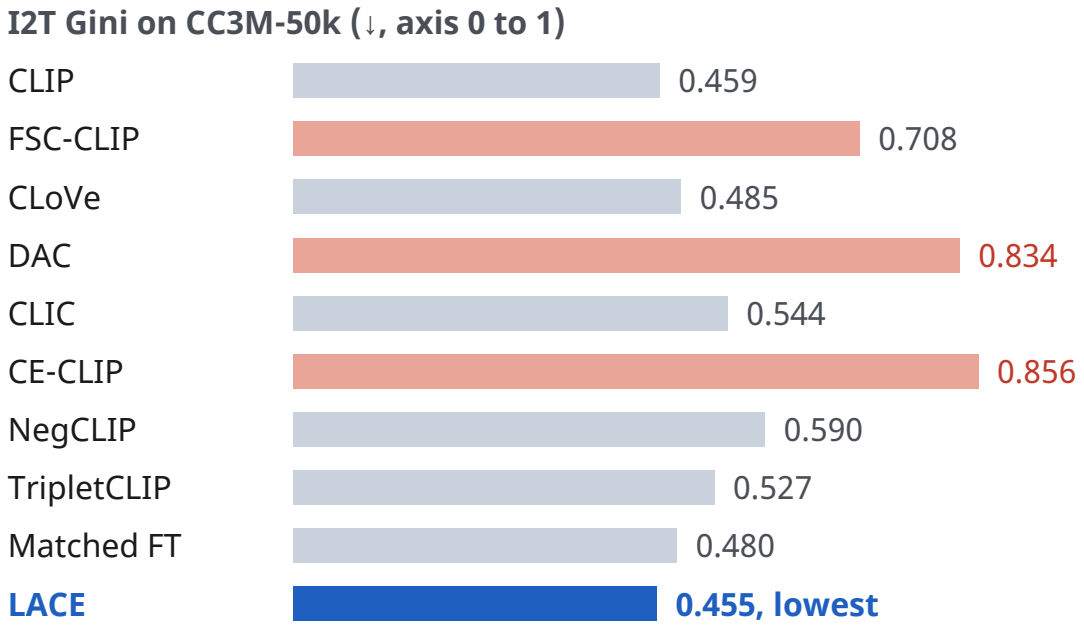
Hubness: Do a Few Points Crowd Out the Rest?

Some points become nearest neighbors for disproportionately many queries.

In-degree count, and its inequality:

$$h_j^T = \left|\{\, i : j \in \mathcal{N}_K^T(i) \,\}\right|, \qquad \mathrm{Gini}(h) = \frac{\sum_{a=1}^{N}\sum_{b=1}^{N}|h_a - h_b|}{2N\sum_{a=1}^{N} h_a}$$

Gini → 1 signals super-hubs that degrade retrieval reliability.

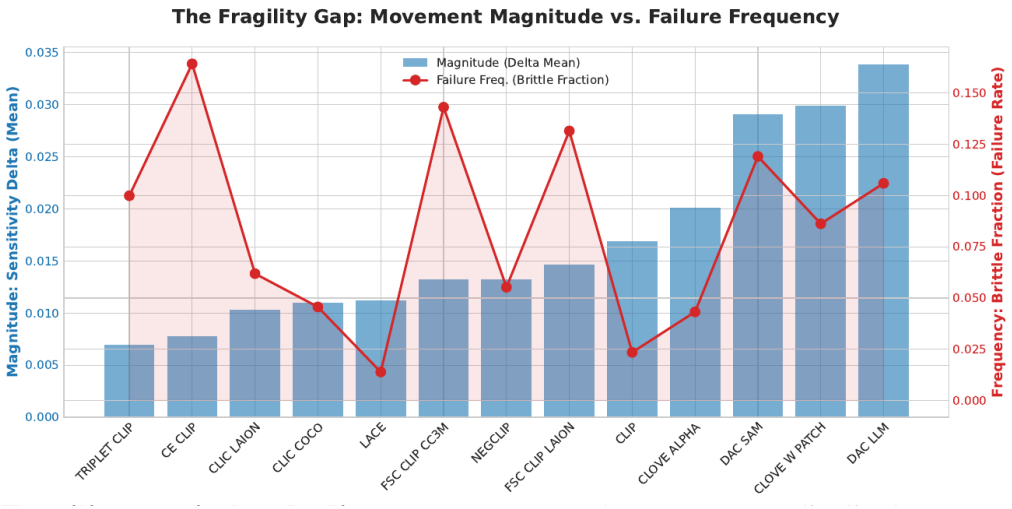


Sec. 4.3 · Eqs. 6–7 · Table 1 · red = Gini above 0.7

Local Edit Geometry and Brittleness

Edit a caption (“cat” → “dog”, “red” → “blue”), then measure the margin: $\Delta_{\text{margin}}(i) = s(\mathbf{v}_i, \mathbf{u}_i) - s(\mathbf{v}_i, \mathbf{u}_i^{\text{edit}})$

$$\text{Brittle}(i) = \mathbb{I}\big[\Delta_{\text{margin}}(i) \le 0\big] \ \vee \ \mathbb{I}\big[I_i \in \mathcal{N}_K^{I,\text{edit}}(i)\big], \qquad \text{BrittleFrac} = \frac{1}{N}\sum_{i=1}^{N} \text{Brittle}(i)$$

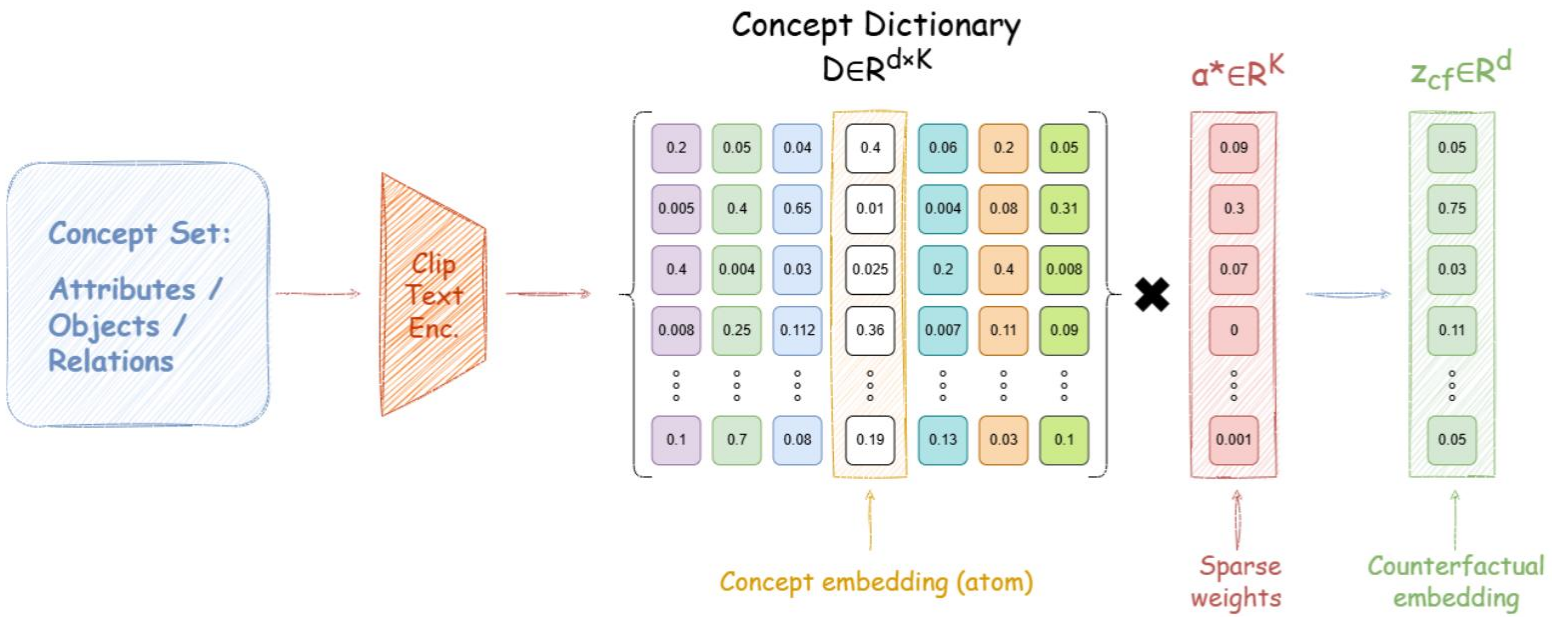


The fragility gap: movement magnitude alone does not predict stability.

Methods with modest displacement can still fail often; LACE stays low on both.

LACE: Counterfactuals in Embedding Space

Edit concepts, not pixels: decompose an embedding into sparse concept atoms, edit a few, reconstruct.



Same operator for images and text: Attenuate Swap Transplant

Sparse Concept Decomposition

Fixed dictionary $D = [\mathbf{d}_1, \ldots, \mathbf{d}_K]$ of concept text embeddings. For $\mathbf{z} \in \{\mathbf{v}, \mathbf{u}\}$:

$$\alpha^\star(\mathbf{z}; D) = \arg\min_{\alpha \in \mathbb{R}^K} \|\mathbf{z} - D\alpha\|_2^2 + \lambda\|\alpha\|_1$$

Editable concepts need a strong code *and* high similarity:

$$\mathcal{K}(\mathbf{z}) = \mathrm{TopP}_k\left(|\alpha_k^\star(\mathbf{z})|\right) \cap \mathrm{TopQ}_k\left(\langle \mathbf{z}, \mathbf{d}_k \rangle\right)$$

This avoids editing spurious atoms. Atoms are bucketed by type so edits stay type-consistent.

FIXED DICTIONARY

14,000

CLIP text atoms, $d = 512$

Objects	**5,200**
Attributes	**5,800**
Relations	**≈ 3,000**

From COCO captions (spaCy) plus Qwen VLM descriptions.

Embedding-Space Counterfactual Synthesis

Replace a source concept $s \in \mathcal{K}(\mathbf{z})$ with a same-type target $t \notin \mathcal{K}(\mathbf{z})$, then reconstruct:

1 · Edit the dictionary

$$\mathbf{d}_k^{(s\to t)} = \begin{cases} \mathbf{d}_t, & k \in \mathcal{K}_s, \\ \mathbf{d}_k, & \text{otherwise} \end{cases}$$

2 · Reuse the sparse code

$$\mathbf{z}^{\mathrm{cf}} = \mathrm{norm}\left(D^{(s\to t)}\,\alpha^{\star}(\mathbf{z})\right)$$

$\mathbf{z} = \mathbf{v}_i \rightarrow$ image-side $\mathbf{v}_{i,k}^{\mathrm{cf}}$

A visual near-miss without any pixel editing.

$\mathbf{z} = \mathbf{u}_i \rightarrow$ text-side $\mathbf{u}_{i,k}^{\mathrm{cf}}$

A caption near-miss from the same operator.

Reusing α⋆ ties the edit size to the source activation: a targeted near-miss that keeps the global context.

LACE: Symmetric Counterfactual Training

Counterfactuals become hard negatives in both retrieval directions:

$$\mathcal{L} = \mathcal{L}_{\text{CLIP}} + \gamma\left(\mathcal{L}^{\text{cf}}_{\text{I2T}} + \mathcal{L}^{\text{cf}}_{\text{T2I}}\right)$$

L^{cf} I2T

Each image vs. text-side counterfactuals $\mathbf{u}^{cf}$.

L^{cf} T2I

Each caption vs. image-side counterfactuals $\mathbf{v}^{cf}$.

L CLIP

Keeps global image–text alignment.

Setting	Value
Backbone and data	OpenCLIP ViT-B/32 + LoRA (r = 16, α = 32), MS-COCO fine-tuning
Defaults	K_cf = 4, γ = 0.5, λ = 0.10, 14k-atom dictionary
Inference	Standard dual encoder: no dictionary lookup or sparse solver

Sec. 5.3 · Eq. 16 · Sec. 6.1, App. D.9

Main Results on Compositional Benchmarks

Best on 7 of 11 metrics, while zero-shot and retrieval also improve over CLIP.

Model	COLA	What'sUp	Wino	SC Add	SC Rep	SC Swap	SC++ Rep	SC++ Swap	IN-v2	Retr.	MMVP
CLIP ViT-B/32	22.86	41.77	16.40	72.68	80.02	62.60	68.68	43.07	55.96	54.67	27.42
DAC-LLMs	18.41	42.57	12.50	93.65	89.33	74.85	53.63	31.95	45.07	47.35	23.71
CLoVe	26.51	41.83	12.67	90.35	88.73	76.15	64.03	40.35	55.28	56.98	22.97
CE-CLIP	19.37	41.73	12.17	91.90	88.53	75.55	65.73	40.10	43.32	61.35	23.70
FSC-CLIP	18.73	40.77	12.58	92.20	86.40	74.05	63.73	39.55	41.01	51.35	20.01
CLIC	22.22	39.79	17.92	86.45	84.73	72.55	76.03	61.60	55.30	57.13	19.28
NegCLIP	20.00	42.43	16.33	87.70	82.92	69.98	70.52	56.55	48.92	62.55	25.93
TripletCLIP	17.46	41.03	15.00	88.34	86.97	69.81	73.68	44.04	41.10	44.24	19.99
LACE (ours)	39.05	43.56	17.42	93.96	86.80	68.17	76.10	48.88	56.66	63.08	35.56
Δ vs CLIP	+16.19	+1.79	+1.02	+21.28	+6.78	+5.57	+7.42	+5.81	+0.70	+8.41	+8.14

Table 2 · accuracy (%), Retr. = mean R@1 · blue = best in column

Ablation Study: Targeted, Two-Sided Edits

Matched init, data, LoRA and schedule; gains need semantic edits on both modalities.



Avg. accuracy (%) over What'sUp, COLA, Winoground, SC, SC++, MMVP. Bars start at zero.

Where COLA's jump comes from

COCO fine-tune alone	+12.70
LACE on top	**+3.49**
Image-disjoint subset	**+2.60**

Tables 24–26 · App. E.5

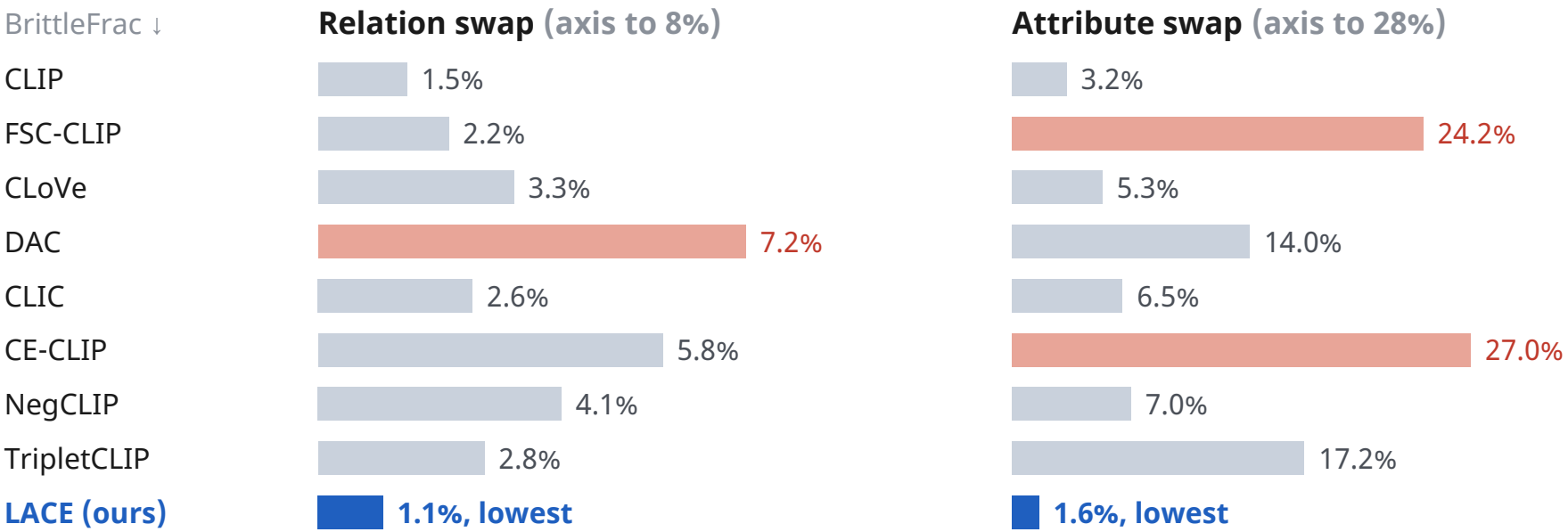
Global Geometry: The Most Reciprocal Space

Text-centric methods raise hubness and cut reciprocity; LACE improves both.

Method	Gini I2T ↓	Top-1% I2T ↓	Gini T2I ↓	Top-1% T2I ↓	Mutual ↑	Edge ↑
CLIP (pretrained)	0.459	7.52	0.493	6.92	50.30	52.18
FSC-CLIP	0.708	11.76	0.536	9.70	30.36	35.95
CLoVe	0.485	5.82	0.426	6.05	45.23	51.17
DAC	0.834	18.81	0.509	7.34	24.07	26.47
CLIC	0.544	7.63	0.473	6.40	38.77	46.45
CE-CLIP	0.856	22.89	0.626	19.63	17.16	23.01
NegCLIP	0.590	9.07	0.469	6.77	35.74	44.48
TripletCLIP	0.527	5.10	0.403	4.57	46.90	50.48
Matched fine-tune	0.480	7.30	0.417	5.40	46.57	52.46
LACE (ours)	0.455	6.96	0.450	6.06	51.41	53.79

Table 1 · CC3M-50k, K = 10, reciprocity in % · blue = best

Local Geometry: Fewest Brittle Failures

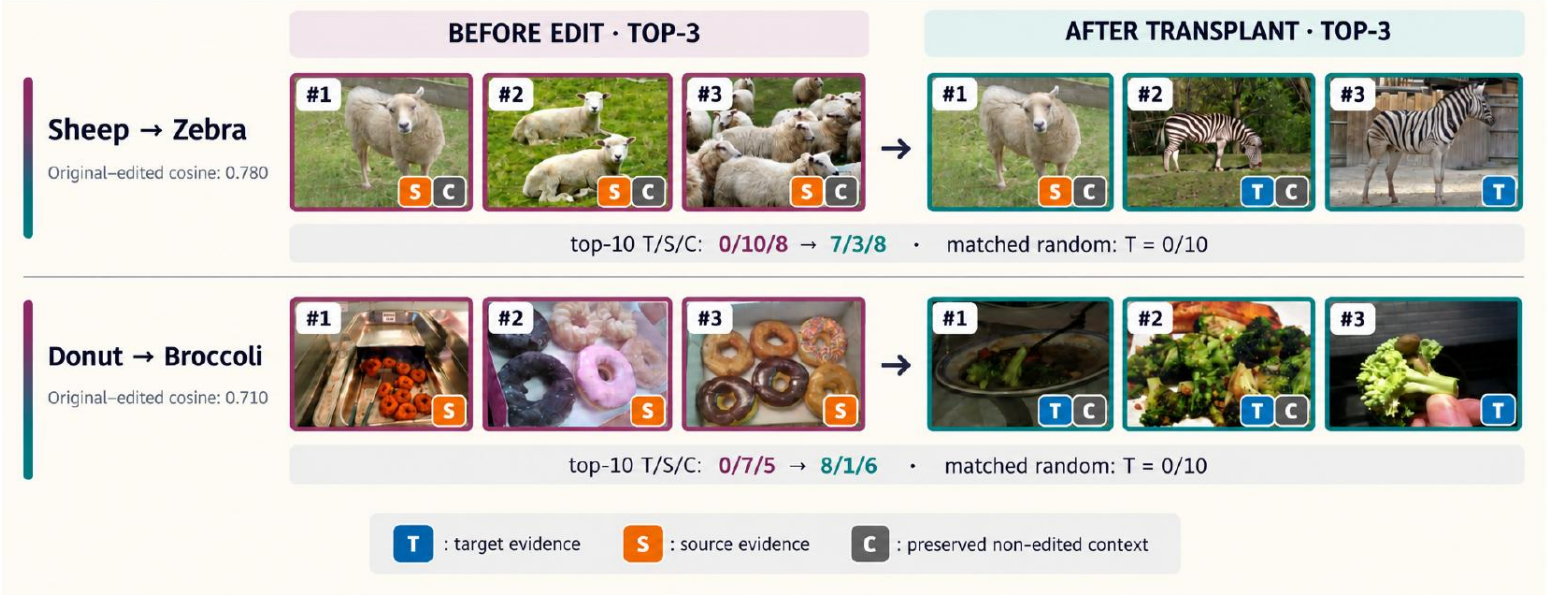


Vs. matched fine-tune: 3.27% → 1.10% (relation) and 4.65% → 1.60% (attribute), ≈ 66% fewer failures.

Table 7 · CC3M-50k, k = 10 · red = highest failures

Qualitative Results: Latent Transplants

Top-3 COCO neighbors before and after an edit: target evidence appears, context is kept.



99.7%

of 606 image edits raise preference for the target

95.9%

of 965 text edits move toward it

Mean shift +0.0122 vs. +0.0002 for matched random edits.

Fig. 3 · App. F.3, G.2

Cross-Objective Validation with SigLIP

Same generator; counterfactuals enter the pairwise sigmoid loss as explicit negative pairs (label −1):

$$\mathcal{L}^{\mathrm{cf}}_{\mathrm{SigLIP}} = \frac{1}{N K_{\mathrm{cf}}} \sum_{i=1}^{N} \sum_{k=1}^{K_{\mathrm{cf}}} \Big[\ell\big(s(\mathbf{v}_i, \mathbf{u}^{\mathrm{cf}}_{i,k}), -1\big) + \ell\big(s(\mathbf{v}^{\mathrm{cf}}_{i,k}, \mathbf{u}_i), -1\big) \Big], \qquad \mathcal{L} = \mathcal{L}_{\mathrm{SigLIP}} + \gamma\, \mathcal{L}^{\mathrm{cf}}_{\mathrm{SigLIP}}$$

Model	What'sUp	COLA	Wino	SC	SC++	MMVP	Avg.
SigLIP ViT-B/16	44.00	37.46	16.67	78.69	57.35	31.85	44.34
SigLIP FT	44.97	34.89	18.00	79.94	64.92	31.85	45.76
SigLIP FT + LACE	47.00	40.90	19.00	83.98	67.95	37.78	49.44

Also not COCO-specific: with GBC training data, LACE lifts the average from 41.83 to 45.76.

Hyperparameter Sensitivity

Stable under reasonable changes; *K*cf = 0 or γ = 0 is the plain fine-tuned baseline. Blue = default.

K_cf	Avg.	γ	Avg.	λ	Avg.	Atoms	Avg.
0	45.04	0.00	45.04	0.01	45.99	500	44.73
1	45.30	0.10	45.10	0.05	46.03	1k	44.98
2	46.10	0.25	45.95	0.10	47.39	2k	45.39
4	47.39	0.50	47.39	0.20	47.21	5k	46.00
8	47.36	1.00	47.00	0.50	46.99	14k	47.39

Gains plateau from 4 counterfactuals per pair; larger dictionaries help (500 → 14k atoms: +2.66).

Table 4 · avg. accuracy (%) over compositional benchmarks

Takeaways

- Text-only hard negatives can win benchmarks while warping the shared space: more hubness, less reciprocity.
- *Sparse concept edits* give cheap image- and text-side counterfactuals, with no pixels or generators.
- *Symmetric counterfactual training* improves binding, reciprocity and local stability.
- Training-only: a plain dual encoder at inference; transfers to SigLIP and GBC.

Project page
harsh-udai.github.io/LACE

Code
github.com/Harsh-Udai/LACE

Thank you

Questions are welcome.

Harsh Udai · ai23resch11003@iith.ac.in

harsh-udai.github.io/LACE

References

(Radford et al., 2021) Learning transferable visual models from natural language supervision. CLIP.

(Yuksekgonul et al., 2023) When and why vision-language models behave like bags-of-words.

(Doveh et al., 2023) Dense and aligned captions (DAC) promote compositional reasoning.

(Castro et al., 2024) CLoVe: Encoding compositional language in contrastive VLMs.

(Oh et al., 2024) Preserving multi-modal capabilities of pre-trained VLMs (FSC-CLIP).

(Zhang et al., 2024) Contrasting intra-modal and ranking cross-modal hard negatives (CE-CLIP).

(Patel et al., 2024) TripletCLIP: Synthetic vision-language negatives.

(Peleg et al., 2025) Advancing compositional awareness in CLIP (CLIC).

(Huang et al., 2025) Visual perturbation and adaptive hard negatives (AHNPL).

(Bhalla et al., 2024) Interpreting CLIP with sparse linear concept embeddings (SpLiCE).

(Ray et al., 2023) COLA: Compositional text-to-image retrieval.

(Kamath et al., 2023) What's "up" with vision-language models?

(Thrush et al., 2022) Winoground: Visio-linguistic compositionality.

(Hsieh et al., 2023) SugarCrepe: Fixing hackable compositionality benchmarks.

(Tong et al., 2024) Eyes wide shut? Visual shortcomings of multimodal LLMs (MMVP).

(Zhai et al., 2023) Sigmoid loss for language image pre-training (SigLIP).