Similarity Is Not Logic

Factored Inference for Dual-Encoder Vision-Language Models

Sultan Alshehri Zhantao Yang Han Zhang Marios Savvides

Carnegie Mellon University

ICML International Conference On Machine Learning

Paper · Code · Benchmark

Dual-Encoder Similarity Is a Bag of Concepts

A single scalar similarity tracks **which** concepts appear, not **how** they combine.

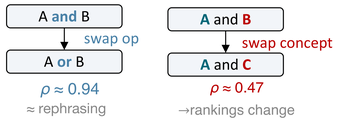
Encoded, but Never Executed

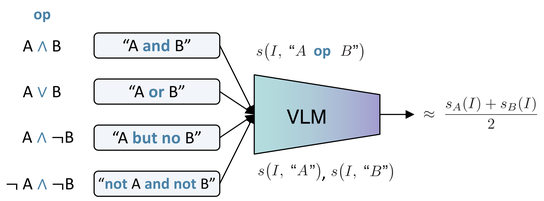
The embedding carries both the concepts and the operator. The dot product acts only on the concepts:

$$g(\text{“}A \text{ and no } B\text{”}) = \underbrace{\mathbf{t}_{\text{span}}}_{\substack{\text{mix of concept embeddings}\\ \text{acts like the } \textit{mean} \text{ and dominates}}} + \underbrace{\mathbf{t}_{\perp}}_{\substack{\text{operator-specific signal}\\ \text{weak or misaligned}}}$$

Encoded. Ranking by the operator part $\mathbf{t}_{\perp}$ alone gets negation 68% correct.

Never executed. The dot product is dominated by $\mathbf{t}_{\text{span}}$, so “no X” still favors images *with* X (12% correct) and swapping “and” $\leftrightarrow$ “or” barely moves rankings (right).





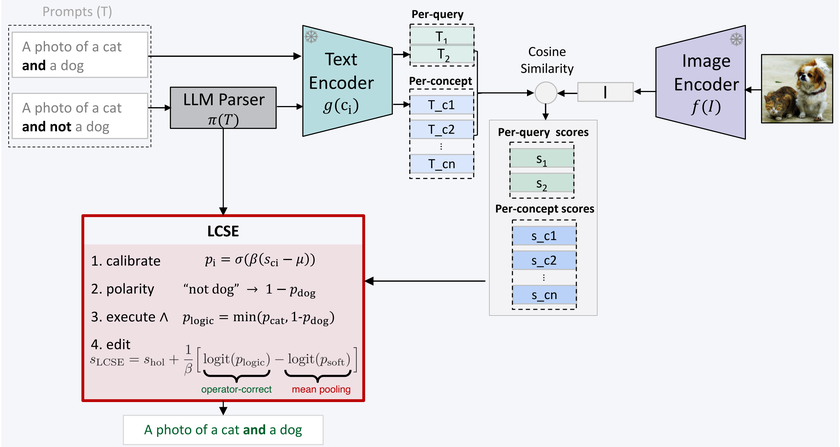
Fine-Tuning Targets the Wrong Bottleneck

Polarity test for “no X”. Do images *without* X outscore images *with* X (AUC $> 50\%$)? **Every tested model**, even negation-finetuned, stays **below chance**. Factored evidence ($1-p$) is correct *by construction*.

Method	AUC
CLIP (Radford et al., 2021)	11.2%
NegCLIP (Yuksekgonul et al., 2022)	9.3%
CoN-CLIP (Singh et al., 2025)	17.2%
CLIP-NegFull (Alhamoud et al., 2025)	12.7%
NegCLIP-NegFull (Alhamoud et al., 2025)	10.9%
NegationCLIP (Park et al., 2025)	19.0%
Factored ($1-p$)	**88.7%**

LCSE: Logic-Constrained Score Editing

Factored inference separates evidence extraction from constraint execution. Score each concept with the *frozen* encoder, execute the operator externally, then *edit* the holistic score:



Training-free · Backbone-agnostic · Retrieval-preserving

FACTOR-Bench

1,695 operator-contrastive caption pairs isolate *operator execution* from concept detection. Concept presence is validated by OWL-ViT on COCO val2017.

Operator split (1,100): NOT · AND · OR · BUT-NOT · NEITHER

Sample. An image with a **car** but no **bicycle**:
- $\wedge$ “car” ✓ vs “car and bicycle” ×
- $\neg$ “no bicycle” ✓ vs “bicycle” ×
- $\vee$ “car or bicycle” ✓ vs “car and bicycle” ×

Same concepts, different operator, so accuracy isolates the *operator*.

Further splits test equivalence (450; De Morgan, double negation, commutativity) and compounds (145; 3–4 concepts, mixed polarity). Balanced polarity and $5+$ templates per operator block text-only shortcuts.

Gain Scales with Evidence Quality

We stratify by the weakest concept's detection AUC. External execution nearly closes the gap where the encoder's evidence is reliable.

Weakest concept AUC	n	Holistic (%)	LCSE (%)	Δ (pp)
High (> 0.90)	417	57.8	**90.4**	$+32.6$
Medium (0.75–0.90)	479	61.4	85.2	$+23.8$
Low (< 0.75)	204	52.0	76.0	$+24.0$
Overall	1,100	58.3	85.5	$+27.2$

Every Operator Improved, Retrieval Preserved

Top-1 retrieval per query (**green** = correct, **red** = wrong). Fine-tuned variants remain unreliable on both operator queries and standard retrieval. LCSE is correct on all five.

85.5%
FACTOR-Bench, CLIP
58.3 → 85.5 **(+27.2 pp)**

90.7%
FACTOR-Bench, SigLIP 2
best absolute accuracy

27 → 65%
NegBench COCO MCQ
SigLIP 2 **(+38 pp)**

$\rho = 0.999$
Retrieval retained
R@5 within 0.1 pp

	FACTOR	NegBench			Retention			
		COCO		VOC	COCO		PASCAL	
Method	Acc↑ (50%)	MCQ↑ (25%)	R@5↑	MCQ↑ (25%)	R@5↑	ρ	R@5↑	ρ
CLIP	58.3	39.3	48.0	38.7	55.3	1.00	82.4	1.00
SigLIP 2 †	65.0	27.2	64.1	27.1	73.5	—	93.2	—
DCSM	53.6	44.8	21.8	44.2	34.6	.63	64.2	.65
CoN-CLIP	63.2	25.7	48.3	39.7	52.3	.79	79.4	.79
NegCLIP	67.2	28.7	64.4	30.5	69.3	.76	89.8	.77
CLIP-NegFull	66.5	54.5	51.9	54.8	54.7	.84	81.3	.85
NegCLIP-NegFull	68.1	56.2	**67.1**	59.6	**69.6**	.80	**90.0**	.81
NegationCLIP	73.2	31.0	65.8	42.1	68.6	.81	89.8	.83
LCSE (CLIP)	85.5	**60.3**	50.8	73.3	55.2	**.999**	82.4	**.999**
LCSE (SigLIP 2) †	90.7	65.2	67.1	87.2	73.5	—	93.1	—
LCSE (NegCLIP)	**86.2**	55.7	65.0	**81.3**	69.3	.76	89.8	.77

	$\neg A$	$\neg A \wedge \neg B$	$A \wedge \neg B$	$A \wedge B$	$A \vee B$
CLIP	58.0	54.4	66.0	68.7	42.0
Best fine-tuned *	77.3	78.4	79.2	**96.0**	62.7
LCSE (CLIP)	**88.0**	**82.4**	**82.8**	84.7	**90.7**

Bold = best among CLIP methods; † = different backbone; ρ = rank correlation vs. CLIP; * column-wise best of the six fine-tuned baselines.

Takeaway

Compositional failure is **interface-level, not representation-level**. The encoder detects the concepts, but the scoring interface fails to combine them. **LCSE executes the logic externally**. With *no retraining*, it improves all seven backbones tested on queries with negation (+3–28 pp).

Project page
`sultanmo.github.io/factored-vlm`
Paper · Code · FACTOR-Bench