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

Beyond Hard Negatives: Grounded Positive Supervision for Compositional CLIP

Harsh Udai[1]* Konda Reddy Mopuri[1] Vineeth N. Balasubramanian[2,3]

[1]Department of AI, IIT Hyderabad [2]Department of CSE, IIT Hyderabad [3]Microsoft Research India

*Corresponding author

Pre-trained VLMs: Successes and Challenges

VLMs like CLIP perform well in:

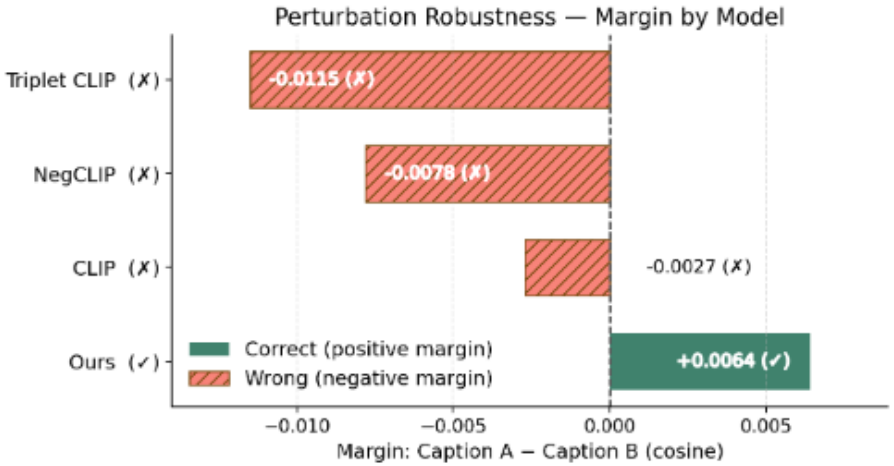
- Zero-shot classification and image–text retrieval.

However, they struggle with *compositional reasoning*: who has which attribute, and who does what to whom.



Caption A:
A thick art book lies to the far left of the coffee table, with the TV remote placed beside it and a stack of coasters set in the center. (Correct)

Caption B:
The TV remote lies to the far left of the coffee table, placed beside a stack of coasters, with a thick art book set in the center.



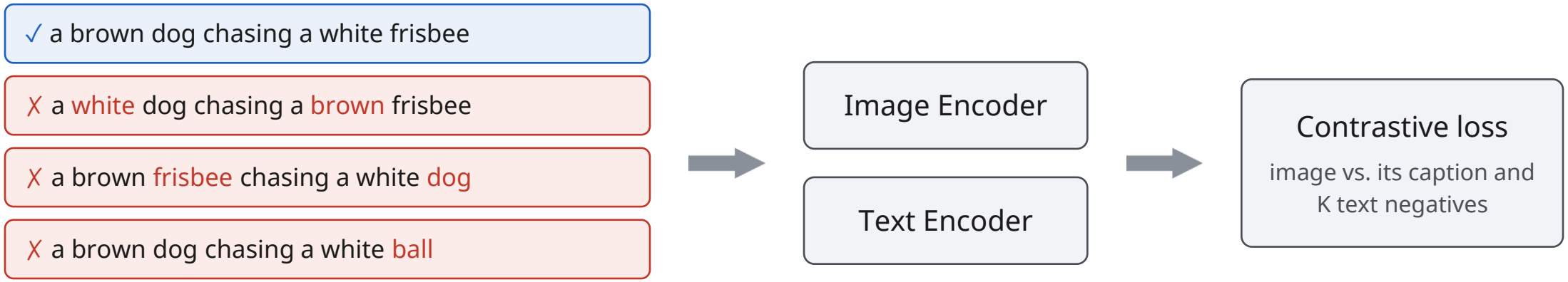
Swapping object positions: CLIP, NegCLIP and TripletCLIP give the wrong caption a higher score (negative margin).

Fig. 1 (top row)

Enhancing Compositional Reasoning in CLIP

Fine-tuning CLIP with *hard negative captions* in the contrastive loss:

- NegCLIP, DAC, CE-CLIP, CLoVe, FSC-CLIP, TripletCLIP, and others.



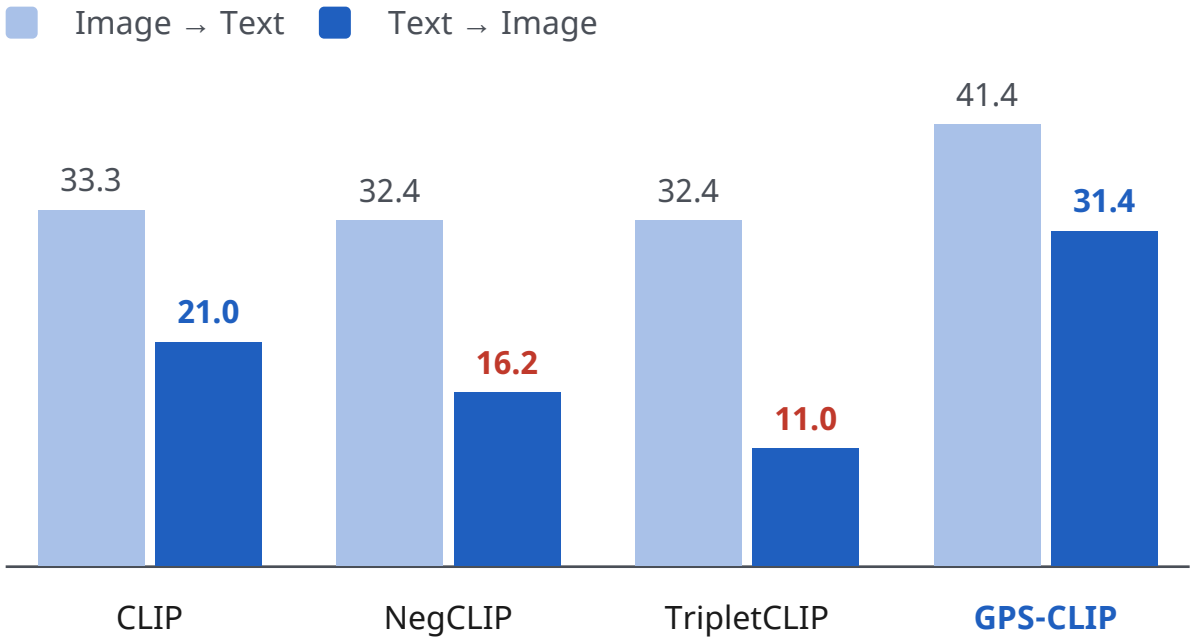
1 caption + K edited captions (hard negatives)

All the fine-grained difficulty sits on the text side.

Yuksekgonul et al., 2023; Doveh et al., 2023; Zhang et al., 2024; Patel et al., 2024

Trade-offs in Fine-tuning CLIP with Hard Negatives

Caption-side hard negatives mainly reshape the text manifold,
but leave the image manifold under-constrained: a modality asymmetry.



COLA multi-object accuracy (%).

Hard-negative methods keep Image → Text, but Text → Image falls below plain CLIP.

GPS-CLIP constrains the image side directly and improves both directions.

Table 14 (COLA multi-object)

GPS-CLIP: Grounded Positive Supervision

Shape the attraction, not just the repulsion: scene graphs define *grounded positives*.

Global
image ↔ caption, standard InfoNCE

Region-level
region ↔ span, IoU-weighted multi-positives

Relation-level
image ↔ ⟨s, r, o⟩ triplets, role-aware

No synthetic negatives, no new modules: plain CLIP at inference.

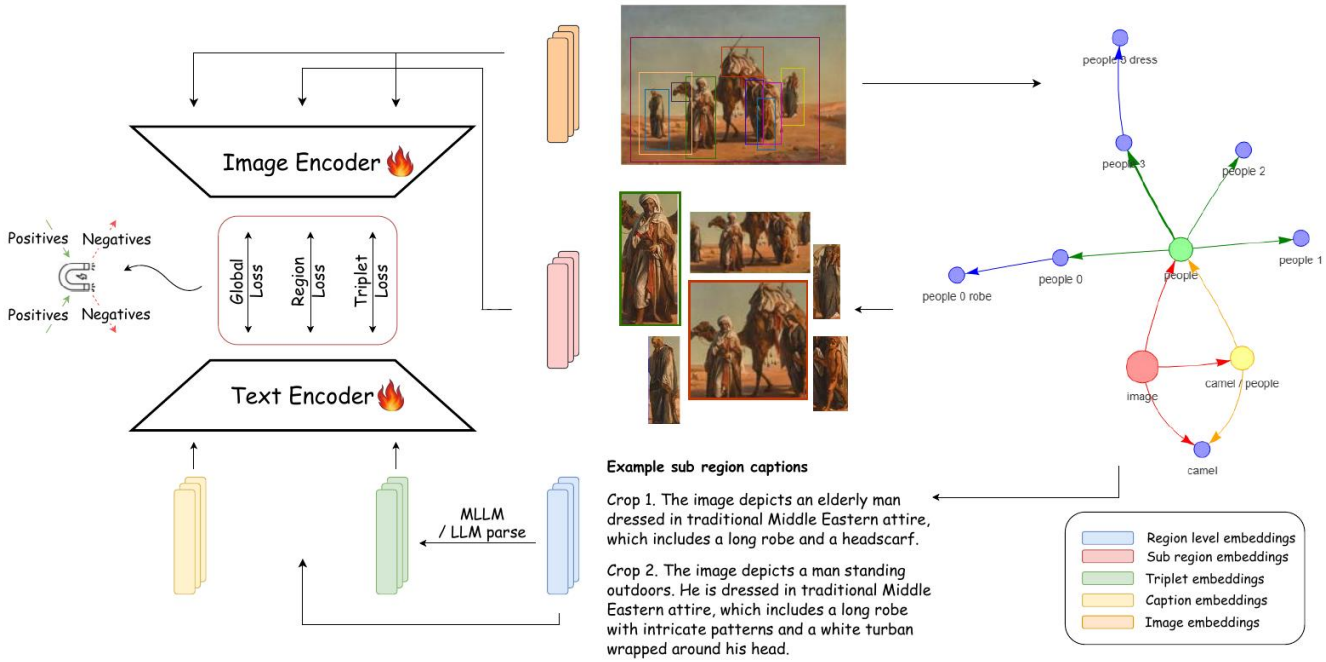


Fig. 2 · Sec. 3

Preliminaries: CLIP-Style Contrastive Learning

A vision encoder $f\theta$ maps image I_i to $\mathbf{v}_i$, a text encoder $g\varphi$ maps caption T_i to $\mathbf{t}_i$.
Both are ℓ_2-normalized; $\text{sim}(\mathbf{v}_i, \mathbf{t}_i) = \mathbf{v}_i^\top \mathbf{t}_i$. Symmetric InfoNCE over N pairs:

$$\mathcal{L}_{\text{global}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\underbrace{\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)}}_{\text{image}\to\text{text}} + \underbrace{\log \frac{\exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_j)/\tau)}}_{\text{text}\to\text{image}}\right]$$

Effective for global alignment, but one pooled vector per image under-specifies *which attribute belongs to which object* and *who does what to whom*.

Overlap in Scene Graphs: Multi-Positives by IoU

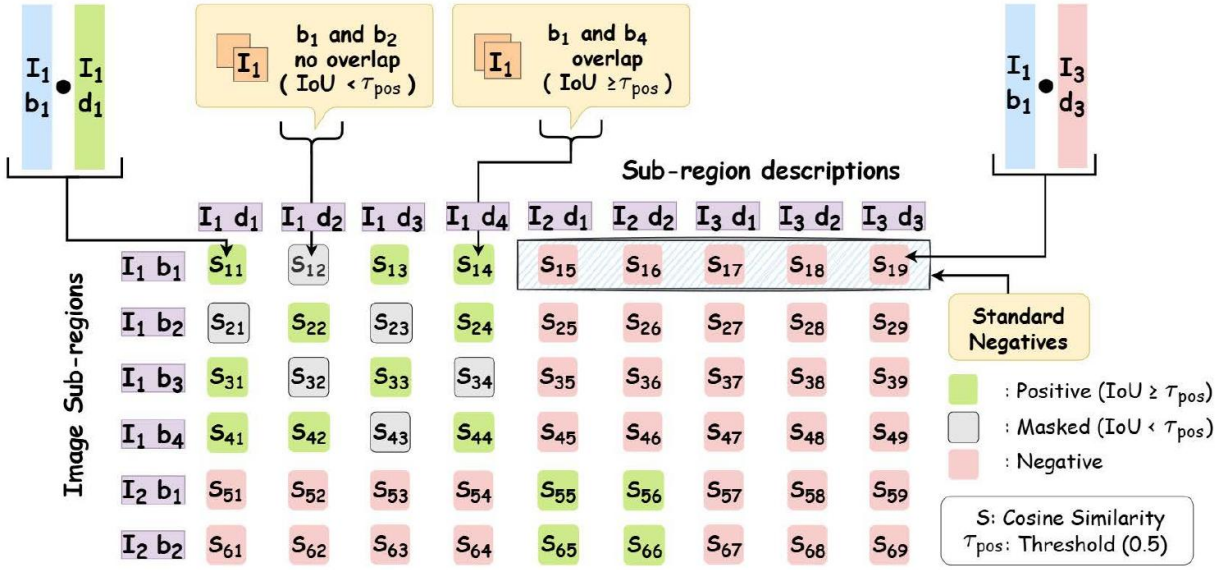
Scene graphs produce *overlapping* regions; treating them as negatives creates false negatives.

Positive set for region b_i in image I:

$$\mathcal{P}(i) = \{\, j \mid \mathrm{IoU}(b_i, b_j) \geq \tau_{\mathrm{pos}} \ \wedge \ b_j \in I \,\}$$

- Positive: same image, IoU ≥ τpos
- **Masked**: same image, IoU < τpos
- Negative: regions from other images

τpos = 0.5. Masked pairs leave the denominator: they still belong to the global scene.



Region-Level Contrastive Alignment

Align region $\mathbf{v}_i^{\text{reg}}$ with caption span $\mathbf{t}_j^{\text{span}}$ using IoU-weighted multi-positives, with $S(\mathbf{v},\mathbf{t}) = \exp\big(\text{sim}(\mathbf{v},\mathbf{t})/\tau\big)$

$$\mathcal{L}_{\text{region}} = -\frac{1}{R}\sum_{i=1}^{R}\log\frac{\sum_{j\in\mathcal{P}(i)} w_{ij}^{r\to t}\, S\big(\mathbf{v}_i^{\text{reg}},\mathbf{t}_j^{\text{span}}\big)}{\sum_{k\in\mathcal{C}_t(i)} S\big(\mathbf{v}_i^{\text{reg}},\mathbf{t}_k^{\text{span}}\big)} \;-\; \frac{1}{M}\sum_{j=1}^{M}\log\frac{\sum_{i\in\mathcal{P}(j)} w_{ji}^{t\to r}\, S\big(\mathbf{t}_j^{\text{span}},\mathbf{v}_i^{\text{reg}}\big)}{\sum_{k\in\mathcal{C}_v(j)} S\big(\mathbf{t}_j^{\text{span}},\mathbf{v}_k^{\text{reg}}\big)}$$

Positive weights
(α = 0.5, γ = 1)

$$w_{ii} = \alpha, \qquad w_{ij} = (1-\alpha)\,\frac{\text{IoU}(b_i,b_j)^{\gamma}}{\sum_{\ell\in P_i}\text{IoU}(b_i,b_\ell)^{\gamma}}, \quad j\in P_i$$

Masked candidate sets

$$\mathcal{C}_t(i) = \{\,k \mid \text{img}(\mathbf{t}_k)\neq\text{img}(\mathbf{v}_i)\,\}\cup\mathcal{P}(i), \qquad \mathcal{C}_v(j) = \{\,k \mid \text{img}(\mathbf{v}_k)\neq\text{img}(\mathbf{t}_j)\,\}\cup\mathcal{P}(j)$$

Sum *inside* the log: positives pool their mass. Masking (+1.46) and IoU weighting (+2.74) are the main contributors.

Relation Awareness: Triplet-Level Contrastive Learning

Captions are parsed into ordered ⟨*s*, *r*, *o*⟩ triplets; each image is aligned with its own: $\mathbf{t}_{sro} = g_\phi(\langle s, r, o \rangle), \qquad \mathbf{v} = f_\theta(I)$

$$\mathcal{L}_{\text{relation}} = -\frac{1}{|\mathcal{R}|} \sum_{(i,m)\in\mathcal{R}} \log \frac{\exp\left(\text{sim}(\mathbf{v}^{(i)}, \mathbf{t}_m^{(i)})/\tau\right)}{\exp\left(\text{sim}(\mathbf{v}^{(i)}, \mathbf{t}_m^{(i)})/\tau\right) + \sum_{(j,n)\in\mathcal{R}\setminus\mathcal{R}_i} \exp\left(\text{sim}(\mathbf{v}^{(i)}, \mathbf{t}_n^{(j)})/\tau\right)}$$

$$-\frac{1}{|\mathcal{R}|} \sum_{(i,m)\in\mathcal{R}} \log \frac{\exp\left(\text{sim}(\mathbf{t}_m^{(i)}, \mathbf{v}^{(i)})/\tau\right)}{\exp\left(\text{sim}(\mathbf{t}_m^{(i)}, \mathbf{v}^{(i)})/\tau\right) + \sum_{j=1,\, j\neq i}^{N} \exp\left(\text{sim}(\mathbf{t}_m^{(i)}, \mathbf{v}^{(j)})/\tau\right)}$$

Log *inside* the sum: each triplet is its own decision, scored only against cross-image candidates.

Role-swap accuracy (ARO-VG, chance 50%)

Global-only 46.7% → + triplet loss 57.1% → GPS-CLIP 57.4%

Sec. 3.6 · Eq. 4 · App. E.10 (Table 27)

GPS-CLIP: Overall Training Objective

Fine-tuning CLIP with three complementary granularities, equal weights (λ = 1):

$$\mathcal{L} = \lambda_{\text{region}} \mathcal{L}_{\text{region}} + \lambda_{\text{relation}} \mathcal{L}_{\text{relation}} + \lambda_{\text{global}} \mathcal{L}_{\text{global}}$$

Setting	Value
Backbone	OpenCLIP ViT-B/32, LoRA rank 16, α = 32, both encoders
Data	GBC-1M filtered to 400k images (≥ 5 regions, ≥ 10 triplets); no MS-COCO
Triplets	Qwen2.5-7B-Instruct, one-time ≈ 3 h on 6 × V100
Optimisation	AdamW, lr 1e-5, 26 epochs, global batch 480, fp16, 4 × V100
Inference	Identical to CLIP: no boxes, graphs, extra parameters or FLOPs

Main Results on Compositional Benchmarks

Best on 8 of 11 metrics with *no* hard negatives, while zero-shot and retrieval also improve.

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-RedCaps	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
GPS-CLIP (ours)	30.64	47.38	18.35	94.78	84.00	70.75	76.17	49.79	58.11	62.92	33.33
Δ vs CLIP	+7.78	+5.61	+1.95	+22.10	+3.98	+8.15	+7.49	+6.72	+2.15	+8.25	+5.91

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

Ablation Study: Data vs. Objective

Region and relation losses are complementary; the gain is not just from better data.



+2.51

from GBC data alone (global loss)

+4.61

from the structured losses on identical data

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

Tables 2, 17, 22

Matched Controls on the Same Data

Identical init, LoRA, batch, schedule and seed; only how structure enters the loss changes.

Model (GBC subset)	What'sUp	COLA	Wino	SC	SC++	MMVP	Avg.
Global loss only	41.90	27.90	17.10	77.60	57.60	28.89	41.83
Naive structured (single-positive)	40.00	30.00	16.58	76.00	60.00	30.00	42.10
NegCLIP-GBC (hard negatives)	43.00	24.44	16.25	81.00	60.21	22.96	41.31
GPS-CLIP	47.38	30.64	18.35	83.29	65.62	33.33	46.44

- Same boxes and triplets with plain InfoNCE: only +0.27; same-image overlaps become conflicting negatives.
- Hard negatives on the same captions help SugarCrepe but lose COLA and MMVP; ImageNetV2 drops to 52.03 (GPS-CLIP 58.11).

Table 28 · App. E.11

Embedding Geometry and Fine-grained Understanding

Grounded positives tighten the image side while keeping text structure intact.

UT-Zappos retrieval ↑

38.2%

Best object–attribute retrieval; next best CLIC 35.0%

Text cross-class ↓

0.783

Least mixing between object classes; NegCLIP 0.833

MMVP-VLM ↑

33.3

Fine-grained visual patterns; CLIP 27.4, +5.9 points

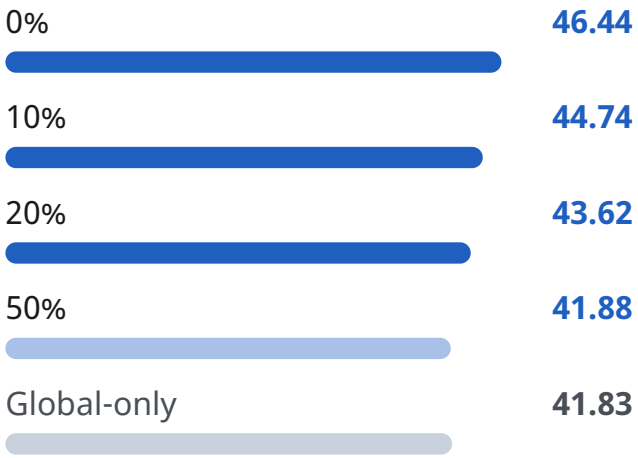
Tables 3, 16, 20

Robustness, Scaling and Loss Generality

Average accuracy (%) over six benchmarks; blue = GPS-CLIP variant.

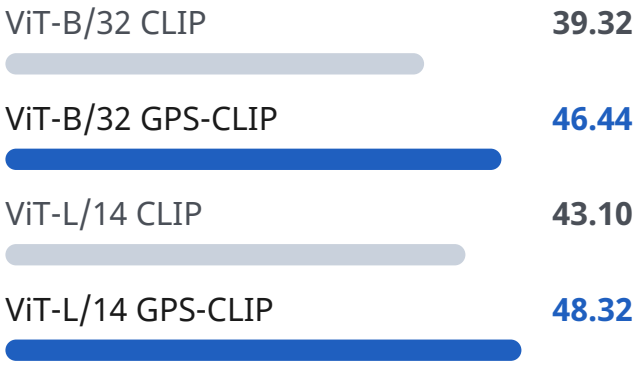
Noisy scene graphs

k% of regions and triplets corrupted



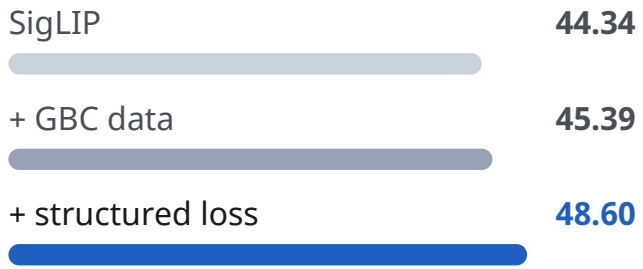
Larger backbone

Same LoRA recipe, light tuning for ViT-L/14



SigLIP sigmoid loss

ViT-B/16; ambiguous same-image pairs ignored



Qualitative Results: Cosine Margins

Margin = sim(correct) – sim(perturbed): green, correct; red hatched, failure.



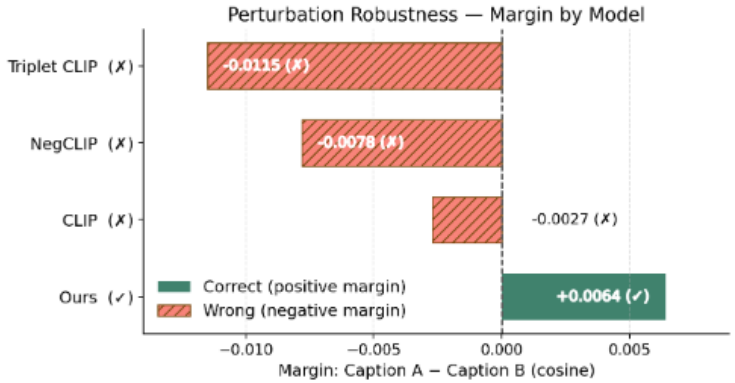
Caption A:
A thick art book lies to the far left of the coffee table, with the TV remote placed beside it and a stack of coasters set in the center. (Correct)

Caption B:
The TV remote lies to the far left of the coffee table, placed beside a stack of coasters, with a thick art book set in the center.



Caption A:
The students are standing on top of the desks, with the teacher seated at the front of the classroom.

Caption B:
The students are standing at the front of the classroom, with the teacher seated behind the desks. (Correct)



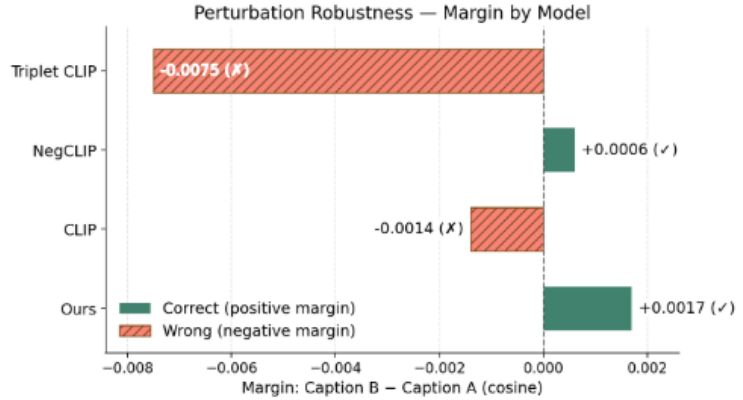


Fig. 1 · more examples in App. G

Takeaways

- Text-side hard negatives sharpen captions but leave the image side under-constrained.
- *Grounded positives* from scene graphs constrain the image side directly, with no synthetic negatives.
- *IoU-weighted multi-positives* with same-image masking are what make structure help.
- Training-only: plain CLIP at inference; gains hold for ViT-L/14 and SigLIP.

Project page

harsh-udai.github.io/GPS-CLIP

Code

github.com/Harsh-Udai/GPS-CLIP

Thank you

Questions are welcome.

Harsh Udai · ai23resch11003@iith.ac.in

harsh-udai.github.io/GPS-CLIP

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. ICLR 2023.

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

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

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

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

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

(Peleg et al., 2025) Advancing compositional awareness in CLIP with efficient fine-tuning (CLIC).

(Hsieh et al., 2025) Graph-based captioning: Interconnecting region captions (GBC).

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

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

(Ray et al., 2023) COLA: A benchmark for compositional text-to-image retrieval.

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

(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).