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, India
[2]Department of CSE, IIT Hyderabad, India [3]Microsoft Research India
ai23resch11003@iith.ac.in, krmopuri@ai.iith.ac.in
vineethnb@cse.iith.ac.in, vineeth.nb@microsoft.com
Project Page Code

Abstract

CLIP-style vision-language models transfer broadly but remain brittle for compositional reasoning, often collapsing object-attribute binding and inter-object relations. We attribute this gap to limited compositional diversity in web-scale data and contrastive objectives that prioritize global alignment over grounded structure. Many fixes rely on caption-edited hard negatives, which can bias learning toward specific perturbation templates and trade off against general alignment. We propose *GPS-CLIP*, which improves compositionality by adding multi-granular positive constraints to standard contrastive learning: **IoU-weighted multi-positive region-span alignment** and **relation-triplet alignment** in the shared embedding space. GPS-CLIP uses existing grounded annotations during fine-tuning, keeps the dual-encoder architecture unchanged, and requires only images and text at inference. Across SugarCrepe, What'sUp, COLA, and Winoground, GPS-CLIP achieves state-of-the-art compositional performance while improving cross-modal retrieval and zero-shot recognition. Beyond compositionality, GPS-CLIP (i) achieves the strongest MMVP-VLM results for fine-grained visual understanding and (ii) yields a more structured embedding space with improved object/attribute separability on UT-Zappos geometric analysis compared to hard-negative variants. Overall, GPS-CLIP offers a streamlined way to improve VLM compositional robustness.

1 Introduction

Humans effortlessly bind objects, attributes, and relations into coherent scene representations that generalize compositionally. From "a brown dog chasing a white frisbee," we immediately infer *who is chasing whom*, can paraphrase the description, and generalize to "a white dog chasing a brown frisbee." This recombination of familiar concepts is a cornerstone of perception, reasoning, and language. By contrast, despite strong coarse alignment and impressive zero-shot performance on classification and retrieval [1–7], modern vision-language models trained on billions of image-text pairs still stumble on attribute binding, role disambiguation, and subtle structural variations.

At the heart of most VLMs lies a *global contrastive function*: paired image-text embeddings are pulled together while mismatched pairs are pushed apart. Popularized by CLIP [1], this objective enables zero-shot recognition, yet operating on a *single global embedding* provides only coarse supervision with no notion of *where* a mismatch occurs. We argue this is compounded by an information imbalance in web-scale data [8], images are information-rich, whereas captions are often *text-sparse*, omitting many attributes and relations visible in the scene [9]. Consequently, models often converge on spurious statistical shortcuts (e.g., object co-occurrence) rather than learning explicit bindings,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

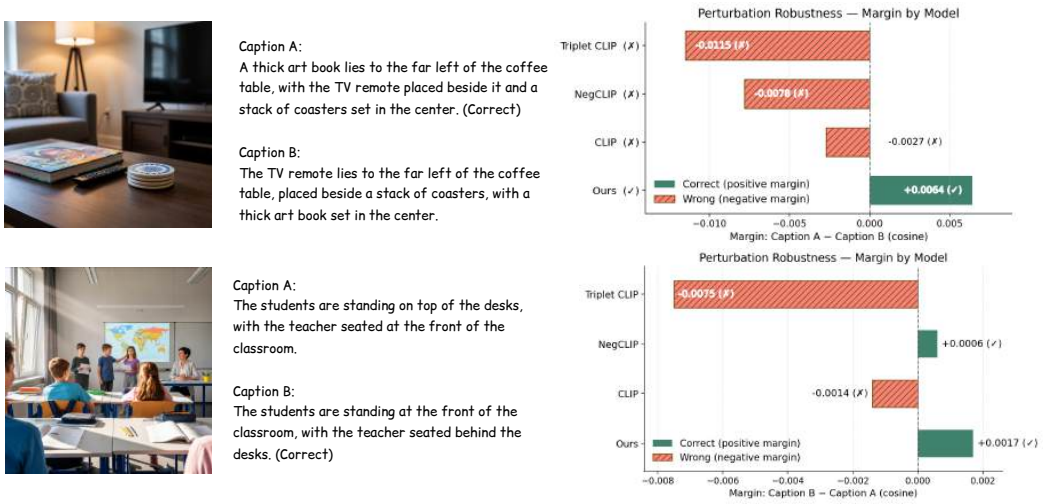


Figure 1: **Compositional robustness of CLIP variants.** Two multi-object scenes with ground-truth and minimally perturbed captions. Cosine margins (m): $m > 0$ (green) correct, $m < 0$ (red) failure. **GPS-CLIP** yields larger margins, especially for subtle attribute/relation changes, indicating stronger fine-grained alignment without synthetic caption corruption (more examples in Appendix G).

rendering them vulnerable to semantic hallucinations when faced with hard negatives. A common response is to fine-tune with caption-side *hard negatives* (HN) [10–14]. However, we argue this primarily reshapes the linguistic manifold, leaving the visual representation under-constrained. This often induces a pathological modality asymmetry, the model becomes sensitive to text perturbations, while the image manifold remains brittle, leading to a pronounced collapse in Text-to-Image (T2I) retrieval [15]. We observe this on multi-object benchmarks such as COLA (Tab. 14), where HN-based methods (e.g., NegCLIP, TripletCLIP) can preserve Image-to-Text (I2T) while suffering a pronounced drop in T2I retrieval. Scaling data and parameters has not eliminated this gap [9], suggesting that standard contrastive learning lacks the **structural inductive bias** needed for role-aware binding. Unlike prior approaches [16, 17] that reshape the text manifold or rely on synthetic negatives, **GPS-CLIP constrains the image manifold through positive-only geometric consistency** for CLIP-style dual encoders. Its central technical idea is an *IoU-weighted multi-positive formulation* that turns scene-graph grounding into graded positive constraints. Instead of using brittle one-to-one matches or synthetic negatives, GPS-CLIP aligns each text span with multiple geometrically compatible regions, weighting positives by spatial overlap. Thus, rather than teaching the model to reject counterfactual captions, GPS-CLIP aligns each image with its own grounded substructures, thereby reshaping the attraction geometry of the joint embedding space. GPS-CLIP introduces three positive-centric mechanisms: (1) ***Local-global synergy:*** IoU-based multi-positive assignment aligns image regions with matched text spans; (2) ***Weighted geometric consistency:*** balances local feature learning with global scene understanding; and (3) ***Relational anchoring:*** a relation-aware contrastive loss over $\langle s, r, o \rangle$ triplets encourages sensitivity to the directional nature of predicates. We control for data effects by fine-tuning a global-only baseline on the same data, and show that adding the structured losses yields an additional $+4.6\%$ average gain (App. E.5). We evaluate GPS-CLIP across compositional benchmarks and fine-grained understanding benchmarks. Our contributions are:

i. **Positive-centric image-side structured alignment.** We introduce a training paradigm that uses scene-graph structure to define grounded positive constraints in the joint embedding space.

ii. **Multi-granular grounding objectives.** We instantiate this principle with an IoU-weighted multi-positive region-span loss and a relation-aware triplet loss over $\langle s, r, o \rangle$ structures, jointly targeting object-attribute binding and directional relation understanding.

iii. **Empirical validation.** GPS-CLIP achieves state-of-the-art results on compositional benchmarks (Table 1), improves fine-grained understanding (MMVP-VLM), zero-shot recognition, and retrieval, and yields improved object-attribute disentanglement in UT-Zappos geometric analysis. Controlled ablations (App. E.5) disentangle data effects from structured learning.

iv. **Scalability and robustness.** GPS-CLIP is a post-pretraining fine-tuning stage that preserves inference, remains robust to scene-graph noise ($\sim$20%, Sec. 4.6), scales to ViT-L (App. E.7), and is loss agnostic and validated on SigLIP-style sigmoid loss (Table 5).

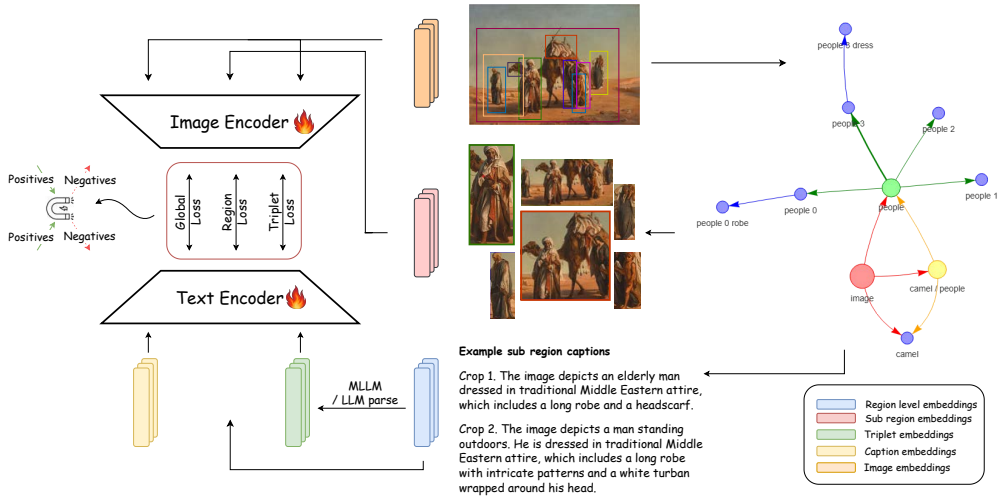


Figure 2: **Structured contrastive at a glance.** An image and caption are decomposed into *regions* and a scene graph. Three complementary losses act jointly: *global* image–caption alignment (λ_{global}), *region-level* IoU-weighted multi-positive alignment (λ_{region}), and *relation-aware* triplet alignment ($\lambda_{\text{relation}}$). Structured outputs: sub-regions with IoU $\geq \tau_{\text{pos}}$ treated as multi-positives (green), and a highlighted $\langle s, r, o \rangle$ path encouraging role/word-order sensitivity.

2 Related Work

Contrastive dual-encoder pretraining at web scale (e.g., CLIP) is the dominant paradigm for vision-language alignment, learning transferable representations by pulling paired image-text embeddings together while pushing apart negatives [1, 2, 18]. It enables strong zero-shot transfer and underpins many multimodal systems [19–21], but its *global-only* objective often misses fine-grained bindings, relations, and word order [10, 22]. Compositional benchmarks such as SugarCrepe, What'sUp, and COLA therefore expose near-chance behavior and "bag-of-words/objects" biases [9, 22–25]. Many extensions improve robustness through data filtering/augmentation, synthetic captions, auxiliary objectives, or hard-negative training via minimal attribute, relation, or word-order edits [13, 26, 27, 11, 28, 10, 29]. However, purely embedding-level hard negatives can degrade global alignment [30], and related work also highlights failures in negation and unimodal composition [31, 32]. Recent methods address these trade-offs through token-patch calibrated hard-negative training (FSC-CLIP) [33], curriculum-style multi-type negatives with intra-modal separation and cross-modal ranking (CE-CLIP) [34], caption densification (DAC) [12], compositional augmentation (CLIC) [14], and weight interpolation for compositionality-generality trade-offs (CLoVe) [35]. We retain the contrastive setup but inject *explicit structure* through IoU-weighted multi-positive region-phrase alignment and relation-aware losses over (s, r, o) triplets. This provides an inductive bias for attribute binding and relational understanding while preserving global alignment. See App. A, B for detailed discussion.

2.1 Structured Representations and Scene-Graph Supervision

A complementary line of work improves compositionality by incorporating **structured representations** (e.g., scene graphs) into the contrastive setup. Structure-CLIP [16] enriches the textual manifold using scene-graph supervision and knowledge-enhanced representations, by relying on hard negatives. MosaiCLIP [36] constructs multi-granularity textual positives and structured hard negatives from scene graphs. Other approaches, such as SGVL [17], incorporate structure via architectural modifications (e.g., auxiliary tokens or heads), requiring structured inputs at inference time. Across these methods, scene graphs are predominantly used to enrich or perturb the *text side*, with limited use for directly constraining image representations. In parallel, region-grounding models (e.g., RegionCLIP [37], GLIP/GLIPv2 [38, 39], Florence-2 [40]) introduce explicit region tokens or pixel-level supervision for fine-grained localization, but at the cost of increased training and inference complexity beyond the standard dual-encoder interface.

Gap in prior formulations. While prior work explores region-text alignment, scene-graph supervision, or grounding, these approaches typically either depart from the standard CLIP-style

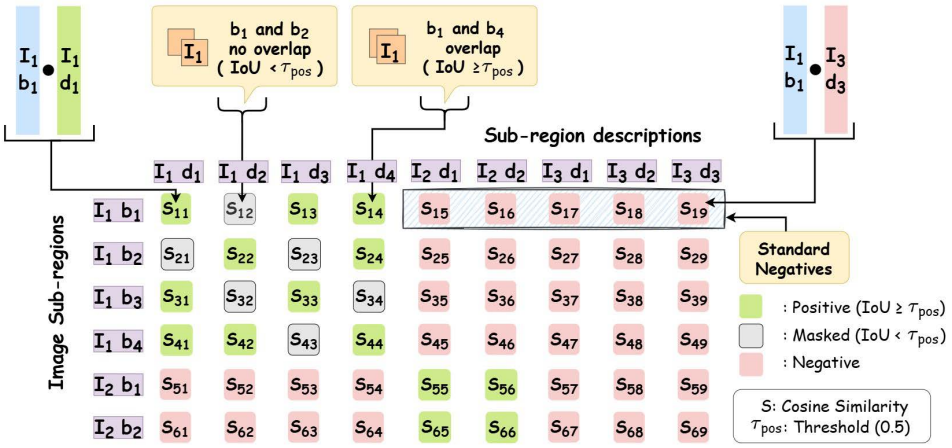


Figure 3: **IoU-weighted region-phrase contrastive loss.** A grid of image sub-regions is compared to phrase descriptions via cosine similarity S. Sub-regions with IoU $\geq \tau_{\text{pos}}$ are treated as multi-positives and *weighted by their IoU*; sub-regions with IoU $< \tau_{\text{pos}}$ have their intra sub-region similarities *masked* (excluded from the positive set), while the remaining pairs act as negatives. This IoU-aware masking and weighting sharpen local grounding and yield more consistent attribute-object bindings.

dual-encoder objective or use structured annotations to refine textual representations rather than to define image-side alignment. As a result, region-level supervision is rarely integrated as *positive alignment signals* within a global contrastive objective. **GPS-CLIP targets precisely this missing regime**: it treats scene-graph structure as a source of grounded positives for shaping the joint embedding geometry, rather than as a source of text-side perturbations, auxiliary graph prediction, or inference-time localization.

GPS-CLIP is orthogonal to both directions. We preserve the standard dual-encoder formulation, introduce no extra architectural modules, and inject scene-graph structure purely through the training objective. Unlike prior scene-graph-based methods that predominantly exploit structure for **text-side augmentation or hard-negative construction**, GPS-CLIP defines **grounded positive alignments** through region-span and relation-level consistency. This positive-centric formulation constrains the joint embedding space without changing the inference-time model architecture or requiring structured inputs at test time. A more detailed analysis is provided in Appendix A.3– A.5 and B.3–B.5.

3 Methodology

We propose GPS-CLIP, a structured contrastive framework (overview in Figure 2) that reformulates image-text alignment via scene-graph decomposition [41]. By bridging sparse captions and dense visual structure, we introduce two objectives targeting VLM compositional failures [22, 25]: (i) a region-level loss aligning sub-regions with matched text spans (Sec. 3.5), and (ii) a relation-aware loss over $\langle s, r, o \rangle$ triplets to explicitly model relational predicates (Sec. 3.6).

3.1 Preliminaries: CLIP-Style Contrastive Learning

CLIP-style VLMs[1] use a dual-encoder architecture: a vision encoder $f_\theta(\cdot)$ maps an image I_i to an embedding $\mathbf{v}_i \in \mathbb{R}^d$ ([CLS] token), and a text encoder $g_\phi(\cdot)$ maps its caption T_i to an embedding $\mathbf{t}_i \in \mathbb{R}^d$ (final [EOS] token). Both are ℓ_2-normalized, and similarity is computed by $\text{sim}(\mathbf{v}_i, \mathbf{t}_i) = \mathbf{v}_i^\top \mathbf{t}_i$. Training uses a symmetric InfoNCE objective over a batch of N image-text pairs, encouraging each image to retrieve its paired caption (*image→text*) and each caption to retrieve its paired image (*text→image*). The loss is defined as follows, where τ is a learnable temperature parameter:

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

3.2 Inter-Modality Structure via Scene Graphs

Web-scale VLM pretraining data is *text-sparse* but *image-rich*: captions often omit attributes/relations that are visually present [9]. This mismatch encourages shortcut solutions. Utilizing scene graphs [41] can reduce this imbalance by decomposing both modalities into fine-grained, aligned units.

Notation. For a caption T, let its scene graph be $\mathcal{G}^{\text{text}} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$ with objects $\mathcal{O}$, attributes $\mathcal{A}$, and relations $\mathcal{R}$; triplets are $\langle s, r, o \rangle \in \mathcal{R}$. For the image I, let $\mathcal{B} = \{b_k\}_{k=1}^{R}$ denote sub-regions (boxes/masks) with associated labels/descriptions consistent with $\mathcal{G}^{\text{text}}$.

Why not directly use Visual Genome (VG) [42]? VG's limited scale, annotation noise, and long-tail sparsity make it insufficient for fine-tuning modern VLMs. We instead use scene-graph annotations as lightweight, source-agnostic supervision.

3.3 Negatives: Roles and Pitfalls (Intra-Image & Cross-Image)

Contrastive learning relies on negatives. Cross-image negatives are standard; however, *intra-image negatives* can be deceptively hard and, if mishandled, push representations to ignore fine-grained correspondences. We therefore explicitly control intra-image negatives by converting overlapping grounded structure into *multi-positives*, rather than treating every non-identical region-span pair as a negative. This changes the attraction structure of contrastive learning and avoids false-negative pressure within the same image.

3.4 Overlap in Scene Graphs: Multi-Positives by IoU

Scene graphs induce overlapping substructures (e.g., subject and subject-attribute regions). To our knowledge, we are the first to formalize *IoU-driven multi-positives* for region-text alignment: for a target region b_i in image I ($b_i \in I$), the positive set is

$$\mathcal{P}(i) = \left\{ j \;\middle|\; \mathrm{IoU}(b_i, b_j) \geq \tau_{\text{pos}} \;\wedge\; b_j \in I \right\}, \tag{2}$$

and $\mathcal{N}(i)$ are the remaining candidates (cross-image) after removing $\mathcal{P}(i)$, with $\mathrm{IoU}(b_i, b_j) \in [0, 1]$ and threshold $\tau_{\text{pos}} \in (0, 1)$. We additionally **mask intra-image negatives** with $\mathrm{IoU}(b_i, b_j) < \tau_{\text{pos}}$ (i.e., same image but non-overlapping/weakly-overlapping regions), since they can be relevant to the *global* scene (coarse alignment) even if not to the target subregion, pushing them apart in a region-level loss would conflict with image-level alignment.

3.5 Region-Level Contrastive Alignment

To move beyond coarse image-text alignment, we propose a *region-level contrastive loss* that aligns image sub-regions with their corresponding caption spans (Fig. 3). Let b_i be a region proposal in image I, with embedding $\mathbf{v}_i^{\text{reg}}$ obtained by encoding the crop of b_i with $f_\theta(\cdot)$ (App. C.2). Let $\mathbf{t}_j^{\text{span}}$ be the embedding of a text span d_j extracted from the scene graph, where each span is a variable-length segment of the caption grounded to a specific sub-region. Unlike prior work restricted to noun phrases or object labels, these spans can be longer descriptive expressions that capture attributes and context. For each region b_i, we define the set of positive spans $\mathcal{P}(i)$ as those that refer to the same object or attribute, including overlapping regions above the IoU threshold. Conversely, for each text span $\mathbf{t}_j^{\text{span}}$, we define its set of positive regions $\mathcal{P}(j)$ in the same manner. We define the similarity score as $S(\mathbf{v}, \mathbf{t}) = \exp(\mathrm{sim}(\mathbf{v}, \mathbf{t})/\tau)$. The IoU-weighted region-level contrastive loss is defined as:

$$\mathcal{L}_{\text{region}} = -\frac{1}{R}\sum_{i=1}^{R} \log \frac{\sum\limits_{j\in\mathcal{P}(i)} w_{ij}^{r\to t} S\big(\mathbf{v}_i^{\text{reg}}, \mathbf{t}_j^{\text{span}}\big)}{\sum\limits_{k\in\mathcal{C}_t(i)} S\big(\mathbf{v}_i^{\text{reg}}, \mathbf{t}_k^{\text{span}}\big)} - \frac{1}{M}\sum_{j=1}^{M} \log \frac{\sum\limits_{i\in\mathcal{P}(j)} w_{ji}^{t\to r} S\big(\mathbf{t}_j^{\text{span}}, \mathbf{v}_i^{\text{reg}}\big)}{\sum\limits_{k\in\mathcal{C}_v(j)} S\big(\mathbf{t}_j^{\text{span}}, \mathbf{v}_k^{\text{reg}}\big)}. \tag{3}$$

Here, $\mathcal{P}(i)$ and $\mathcal{P}(j)$ are multi-positive sets based on IoU and semantic unit overlap (see Sec. 3.4). The directional weights $w_{ij}^{r\to t}$ and $w_{ji}^{t\to r}$ are normalized over the corresponding positive sets and assign greater contribution to strongly overlapping regions. Their exact construction, edge cases, and algorithm are provided in App. C.2. The *masked candidate sets* in the denominator are defined as:

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

Here $\mathrm{img}(\cdot)$ returns the *image ID* associated with a given item: $\mathrm{img}(\mathbf{v}_i^{\text{reg}})$ is the ID of the image containing region i, and $\mathrm{img}(\mathbf{t}_k^{\text{span}})$ is the ID of the caption/span k belongs to. This formulation

excludes the same image but non-positive region/span pairs from the denominator (i.e., those with IoU $< \tau_{\text{pos}}$ or mismatched semantic units), while still including cross-image candidates and in-image positives. This prevents the region-level loss from conflicting with global image-text alignment.

Why is the log *outside the summation (second order summation)?* With multiple valid positives, summing in the numerator aggregates their mass. The outer log then scores the *total positive probability*, so positives reinforce rather than compete.

This mechanism offers two key benefits. First, by treating IoU-overlapping regions as valid positives, the model avoids penalizing near-duplicate substructures that would otherwise be misclassified as negatives. Second, aligning sub-regions with text spans explicitly enforces attribute-object binding at the *part level*, mitigating CLIP's bias towards large or central objects [8]. In effect, $\mathcal{L}_{\text{region}}$ encourages attention to be distributed across finer details, improving compositional grounding. A sequential component ablation identifies same-image masking and IoU weighting as the principal contributors ($+1.46$ and $+2.74$ points, respectively, detailed in Appendix E.9).

3.6 Relation Awareness: Triplet-Level Contrastive Learning

Relational structure is challenging for VLMs because web captions rarely specify spatial and semantic relations. Prior work often uses dense captions or scene-graph-aware *hard negatives* [16], which can be noisy or trivial. Instead, we propose a *relation-aware* contrastive loss over ordered $\langle s, r, o \rangle$ triplets, directly teaching *who does what to whom* beyond global or region-level alignment.

Triplet representations and setup. Given a textual triplet $\langle s, r, o \rangle$, we form a textual embedding $\mathbf{t}_{sro} = g_\phi(\langle s, r, o \rangle)$ (e.g., span-composed encoding of subject, relation, and object). On the visual side, we use the full-image embedding $\mathbf{v} = f_\theta(I)$ since relations are scene-level. Each image is aligned against all triplets extracted from its caption. Let $\mathcal{R} = \{(i, m)\}$ be all (image, triplet) pairs in the batch. Let $\mathcal{R}_i = \{(i, m) \in \mathcal{R}\}$ denote the triplets from image i. Let N be the number of images.

$$\begin{aligned}\mathcal{L}_{\text{relation}} = &-\frac{1}{|\mathcal{R}|} \sum_{(i,m)\in\mathcal{R}} \log \frac{\exp\big(\operatorname{sim}(\mathbf{v}^{(i)}, \mathbf{t}_m^{(i)})/\tau\big)}{\exp\big(\operatorname{sim}(\mathbf{v}^{(i)}, \mathbf{t}_m^{(i)})/\tau\big) + \sum_{(j,n)\in\mathcal{R}\setminus\mathcal{R}_i} \exp\big(\operatorname{sim}(\mathbf{v}^{(i)}, \mathbf{t}_n^{(j)})/\tau\big)} \\ &-\frac{1}{|\mathcal{R}|} \sum_{(i,m)\in\mathcal{R}} \log \frac{\exp\big(\operatorname{sim}(\mathbf{t}_m^{(i)}, \mathbf{v}^{(i)})/\tau\big)}{\exp\big(\operatorname{sim}(\mathbf{t}_m^{(i)}, \mathbf{v}^{(i)})/\tau\big) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \exp\big(\operatorname{sim}(\mathbf{t}_m^{(i)}, \mathbf{v}^{(j)})/\tau\big)}.\end{aligned} \tag{4}$$

Why is the log *inside the summation?* It makes each triplet a separate contrastive decision. With intra-image masking, every triplet scores against *cross-image* candidates only, so we average per-triplet log-likelihoods without penalizing same-image triplets.

This symmetric mechanism ensures that (i) each image embedding retrieves its corresponding relational triplets (*image* $\rightarrow$ *triplet*), and (ii) each triplet embedding retrieves its source image (*triplet* $\rightarrow$ *image*). Modeling triplets at the contrastive level makes the model sensitive to structured relational distinctions such as subject-object role swaps, relational directionality, and predicate binding, known failure modes of CLIP-style models. We do not treat $\langle s, r, o \rangle$ triplets as a complete representation of compositional semantics, rather they provide a tractable relational inductive bias targeting a specific subset of compositional errors. This objective complements the region-level term (local subregion-span alignment) and the global term (overall image-text consistency). Together, the three objectives operate at complementary granularities–global coherence, local grounding, and relational structure-without imposing a fixed symbolic representation. Aggregating the triplet loss over all triplets per image further promotes consistent relational structuring across the scene. We consequently interpret it as an ordered relational inductive bias rather than a predicate-specific grounding mechanism. Direct role-swap and spatial-direction diagnostics are reported in Appendix E.10.

3.7 Structured Fine-tuning Objective

Our final training criterion combines the three terms:

$$\mathcal{L} = \lambda_{\text{region}}\mathcal{L}_{\text{region}} + \lambda_{\text{relation}}\mathcal{L}_{\text{relation}} + \lambda_{\text{global}}\mathcal{L}_{\text{global}}. \tag{5}$$

The framework is *parameter-efficient*, plug-and-play over CLIP fine-tuning, and agnostic to the specific source of scene-graph annotations. While we use InfoNCE for clarity, the same positive constraints can be used with sigmoid-based objectives such as SigLIP [6], as validated in Table 5.

4 Experiments

Our experimental evaluation is structured around three key questions:

- **Q1:** Does purely positive structural supervision improve compositional reasoning?
- **Q2:** Do we achieve improved embedding geometry without relying on hard-negative baselines?
- **Q3:** Do GPS-CLIP's gains extend to fine-grained understanding and SigLIP-style training?

4.1 Experimental Setup

Dataset. We use the **Graph-Based Captioning (GBC)** dataset [41], which provides scene-graph supervision through object regions, bounding boxes, attributes, and relations. Unlike CC3M/12M, which typically provide a single caption per image and suffer from image-text information imbalance [8], GBC offers grounded structured annotations that better expose fine-grained visual content. Importantly, these annotations are not manually curated per image, they are generated automatically at scale using open-vocabulary detectors and multimodal LLMs, making GBC a scalable "silver-standard" source rather than costly human annotation. More broadly, similar MLLM/open-vocabulary detection pipelines could generate silver-standard scene-graph supervision for larger image-text corpora, making GPS-CLIP fine-tuning extensible beyond GBC without requiring manual annotation.

Training setup. We position GPS-CLIP as a post-pretraining compositional alignment stage that injects structured supervision while preserving the CLIP architecture and inference pipeline. We used ViT-B/32 OpenCLIP model [43] pretrained on LAION-2B, and fine-tune LoRA adapters on the GBC-1M split. We extract $\langle s, r, o \rangle$ triplets from the structured annotations and refine them via LLM-based parsing, this only converts existing captions into triplets and introduces no external visual supervision or additional image-text pairs. We filter to images with $\geq$5 subregions and $\geq$10 triplets, yielding 400k training samples. The same structured annotations are used for both our proposed objectives and controlled baselines that remove our inductive biases. All baselines are fine-tuned under identical conditions, with an effective batch size of 480. Additional details in App. C.

Triplet extraction. We extract $\langle s, r, o \rangle$ triplets from captions using Qwen2.5-7B-Instruct. Prompting and implementation details are provided in App. C.3; statistics are reported in App. C.5.

Baselines. We compare against CLIP-style compositional training variants: (i) CLIP zero-shot [1], (ii) a global-only CLIP fine-tune baseline, and (iii) NegCLIP [10]. We also include strong compositional baselines from prior work, e.g., DAC, CLoVe, CE-CLIP, FSC-CLIP, TripletCLIP and CLIC [12, 35, 34, 33, 44, 14]. GPS-CLIP also outperforms a matched NegCLIP baseline trained on the same GBC subset across all six benchmarks, along with detailed discussion in Appendix E.11.

Benchmarks. We evaluate compositionality on COLA [24], What'sUp [9], Winoground [22], and SugarCrepe/SC++ [25]. We further report ImageNetV2 top-1 score, retrieval (Recall@1, avg over COCO and Flickr30k), fine-grained understanding on MMVP-VLM, UT-Zappos for geometric representation analysis. Since SC/SC++ and parts of What'sUp are COCO-derived, we avoid MS-COCO supervision during training to reduce potential *in-domain* overlap and prioritize comparisons with similarly constrained baselines. For prior scene-graph methods [16, 17] whose training or inference assumptions differ substantially from ours, we provide a technical comparison in Sec. 2.1 and App. A.3, and isolate the structured-supervision confound through matched GBC controls.

4.2 Compositional Evaluation

(Q1) Table 1 reports scores on **COLA and What'sUp.** GPS-CLIP achieves the best *average* accuracy on both benchmarks, improving over CLIP by +7.78 on COLA (22.86 $\rightarrow$ 30.64) and +5.61 on What'sUp (41.77 $\rightarrow$ 47.38). Since these benchmarks emphasize multi-object binding and relation/attribute flips, the consistent lift supports that IoU-weighted region supervision strengthens fine-grained correspondences beyond dataset-specific perturbations. **SugarCrepe (SC and SC++).** Across SC/SC++, GPS-CLIP achieves its largest gain on *Add*, reaching 94.78 on SC-Add (+22.10 over

Table 1: **Overall benchmark summary (avg. scores).** We report accuracies (%) on COLA, What'sUp, Winoground, SugarCrepe (Add/Replace/Swap avg), SugarCrepe++ (Replace/Swap avg), ImageNetV2 Top-1, MMVP-VLM Avg, and **Ret Avg** (mean Recall@1 over I2T/T2I, averaged across MS-COCO and Flickr30k). Best/second-best per column are **bold**/underlined. Each benchmark containing multiple subtasks, we report detailed results in App. D.1-D.7.

Model	COLA	What'sUp	Wino	SC			SC++		IN-v2	Ret	MMVP
	Avg	Avg	Avg	Add	Replace	Swap	Replace	Swap	Top-1	Avg	Avg
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 (w/o patching)	26.51	41.83	12.67	90.35	88.73	**76.15**	64.03	40.35	46.05	56.98	22.97
CE-CLIP	19.68	41.73	12.17	91.90	88.53	75.55	65.73	40.10	43.32	61.35	23.70
FSC-CLIP (CC3M)	17.46	40.97	13.33	90.18	85.06	71.12	64.76	41.31	48.06	56.90	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	16.42	88.34	86.97	69.81	73.68	44.04	48.37	53.56	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 *(Ours − CLIP)*	+7.78	+5.61	+1.95	+22.10	+3.98	+8.15	+7.49	+6.72	+2.15	+8.25	+5.91

CLIP), indicating better grounding of newly introduced concepts. It also improves harder edit types, including SC Replace/Swap (+3.98/+8.15) and SC++ Replace/Swap (+7.49/+6.72), and obtains the best SC++ Replace score (76.17). While hard-negative baselines can perform better on *Swap*, likely because they are explicitly trained on swap-style perturbations, GPS-CLIP shows a more balanced profile, it improves Replace/Swap without sacrificing COLA or What'sUp performance. Together with gains on ImageNetV2 zero-shot classification, cross-modal retrieval, and MMVP-VLM in Tab. 1, these results indicate that region- and relation-level positive supervision improves compositional reasoning while improving zero-shot, retrieval, and fine-grained multimodal understanding.

Does structured supervision alone suffice? To separate the effect of scene-graph access from our objective design, we compare against a controlled baseline that uses the same GBC boxes and triplets as GPS-CLIP, but trains region-span and image-triplet pairs independently with standard single-positive InfoNCE, without IoU-based multi-positive grouping or intra-image masking (App. E.11). As shown in Table 28, this naive structured baseline provides only a marginal gain over CLIP-GBC (global) (+0.27 avg) and even degrades several benchmarks, including What'sUp and Winoground. This suggests that simply adding structured supervision can introduce conflicting gradients, since same-image regions with partial semantic overlap are incorrectly treated as negatives. By contrast, GPS-CLIP masks intra-image negatives and promotes IoU-overlapping regions as positives, achieving a consistent +4.34 gain under identical supervision. Thus, the gains come from *how* structure is integrated into the objective, not merely from access to structured annotations. This baseline is more diagnostic than directly re-implementing prior scene-graph methods [16, 17] under mismatched assumptions, because it holds fixed the structured annotations, backbone, training data, and compute while removing only the GPS-CLIP specific objective design choices.

4.3 Ablation Studies

We evaluate the impact of individual loss terms, encoder freezing and weight interpolation (App. E.1-E.3, E.5).

Ablation overview. We structure ablations around three questions: (i) *Data vs. objective*: do gains come from stronger supervision or from the proposed objectives? (ii) *Component vs. recombination*: are gains due to individual losses or their structured integration? (iii) *Design choice*: are weighting and masking necessary? Table 2 summarizes the main trends; detailed comparisons, decompositions, and design-choice studies are in App. A.2, B and E.5.

Table 2: **Disentangling data, components, and design choices (summary).** We report the overall average accuracy (%) across compositional benchmarks and MMVP-VLM; full results in App. E.5.

Model	Region	Triplet	Avg.
CLIP Pretrained			39.32
CLIP-GBC (global)			41.83
CLIP-GBC + Reg.	✓		44.31
CLIP-GBC + Tri.		✓	44.23
GPS-CLIP (Ours)	✓	✓	**46.44**

Table 3: **Multimodal analysis on UT-Zappos-50k.** We report retrieval *Acc* and intra-/cross-class cosine similarities for image and text embeddings (object/attribute groups). Lower *Cross* (↓) indicates less cross-class mixing. **Bold** marks the best value among methods with **Acc** $\geq 30\%$.

Method	Acc (%)	Image (intra)		Image (cross)		Text (intra)		Text (cross)	
		Obj	Attr	Obj↓	Attr↓	Obj	Attr	Obj↓	Attr↓
DAC-LLMs	34.76	0.721	0.764	**0.682**	**0.689**	0.923	1.000	0.845	0.862
FSC-CLIP (CC3M)	32.34	0.801	0.831	0.769	0.776	0.930	1.000	0.843	0.864
CLIC-RedCaps	34.97	0.797	0.826	0.769	0.775	0.937	1.000	0.858	0.877
NegCLIP	34.40	0.816	0.843	0.785	0.792	0.927	1.000	<u>0.833</u>	<u>0.856</u>
TripletCLIP	31.51	0.871	0.889	0.849	0.854	0.955	1.000	0.884	0.903
GPS-CLIP (Ours)	**38.21**	0.763	0.794	<u>0.716</u>	<u>0.728</u>	0.885	1.000	**0.783**	**0.803**

Table 4: **Robustness to scene-graph noise.** We randomly corrupt $k\%$ of region spans and triplet relations in the input scene graphs during training of GPS-CLIP and report average accuracies (%) over compositional benchmarks and the fine-grained MMVP-VLM [46] diagnostic.

Model	WhatsUp	COLA	Winoground	SC	SC++	MMVP	Avg.
GPS-CLIP (0% noise)	**47.38**	**30.64**	**18.35**	**83.29**	**65.62**	**33.33**	**46.44**
GPS-CLIP (10% noise)	<u>45.90</u>	<u>30.00</u>	<u>17.80</u>	<u>82.93</u>	<u>62.90</u>	<u>28.89</u>	<u>44.74</u>
GPS-CLIP (20% noise)	45.20	29.10	16.60	81.32	61.35	28.15	43.62

4.4 Representation Analysis

(Q2) We use **UT-Zappos** as a controlled probe of embedding geometry because it contains fine-grained object-attribute variations. We measure intra-/cross-class cosine similarity within each modality, per-modality variance (App. E.4, Tab. 21), and cross-modal retrieval to assess object-attribute disentanglement and fine-grained alignment. As shown in Tab. 3, GPS-CLIP achieves the best retrieval accuracy, the lowest text-side cross-class similarity (Obj↓, Attr↓), and improved image-side separability relative to hard-negative baselines. Variance analysis further suggests more balanced object/attribute structure, although text-attribute variance remains compressed across models. **Object-Attribute Bias.** Attributes remain harder to align than objects, consistent with object-centric VLM biases [45] and the rarity of attributes appearing independently in training data. Unlike hard-negative training, which can overemphasize dominant cues, GPS-CLIP injects controlled object-attribute variation through structured supervision, improving separability without sacrificing fine-grained alignment. This supports our central geometric claim: grounded positive constraints reshape the joint embedding space by reducing object-attribute mixing while preserving cross-modal alignment. A qualitative PCA projection in App. E.4 provides a visual sanity check of this trend.

4.5 Fine-grained understanding

(Q3) We evaluate fine-grained compositional understanding on MMVP-VLM [46], a multiple-choice benchmark over natural images probing attributes such as orientation, state, etc. As shown in Table 16 and summarized in Table 1, GPS-CLIP achieves the best mean accuracy (**33.3**%), improving over CLIP by +5.9 points (27.4% → 33.3%; ∼21% relative) and outperforming all other variants. These gains require *no explicit hard-negative mining*, suggesting that region- and relation-aware *positive* supervision transfers beyond retrieval-style benchmarks to fine-grained visual understanding.

4.6 Robustness and Practicality Analysis

Robustness to noisy scene graphs. We corrupt $k\%$ of region-span pairs and $\langle s, r, o \rangle$ relations during training. As shown in Table 4, performance degrades smoothly with noise yet remains above CLIP across benchmarks, suggesting that structured supervision acts as a soft inductive bias and does not require perfectly accurate scene graphs. Robustness is non-uniform: MMVP, whose gains depend most on accurate structured supervision, is the most noise-sensitive benchmark (App. D.7). Detailed discussion in App. E.6.

Scalability and efficiency. GPS-CLIP scales as an inference-preserving fine-tuning method, structured annotations are used only during training, and inference remains identical to the underlying dual

Table 5: **SigLIP compatibility.** GPS-CLIP structured supervision improves SigLIP across compositional benchmarks and MMVP-VLM, showing that the alignment principle is not InfoNCE-specific.

Model (ViT-B/16)	SC	SC++	Winoground	MMVP	COLA	What'sUp
SigLIP	78.69	57.35	16.67	31.85	37.46	44.00
SigLIP + GBC (global)	81.00	56.00	18.00	33.33	39.00	45.00
SigLIP + GBC + Structured Loss	**84.40**	**63.00**	**18.17**	**38.52**	**41.00**	**46.50**

encoder, with no additional latency or FLOPs (App. C.1). Its practical scalability is supported along three dimensions. First, it is robust to imperfect structure: corrupting 10% and 20% of region-span and relation annotations degrades average accuracy smoothly from 46.44 to 44.74 and 43.62, while remaining above the CLIP baseline of 39.3 (Table 4). Second, its gains persist across backbone capacity, improving over CLIP when scaling from ViT-B/32 to ViT-L/14 (App. E.7). Third, because the overhead is confined to training, GPS-CLIP remains compatible with automatically generated silver-standard scene-graph supervision. We discuss broader paths for large-scale scaling in App. F.

Generalization beyond CLIP-style contrastive loss. GPS-CLIP is not tied to InfoNCE. Replacing it with a SigLIP-style [6] sigmoid loss yields comparable gains (Table 5), IoU-matched region-span pairs are positives, cross-image mismatches are negatives, and ambiguous same-image non-positives are ignored. This supports GPS-CLIP as a drop-in compositional alignment stage for modern dual-encoder VLMs beyond standard CLIP. Detailed design discussion in App. E.8.

Additional discussions. Complete training and implementation details are provided in App. C. Additional analyses include qualitative examples (Fig. 4; App. G), controlled structured baselines (App. E.11), multi-positive analysis (App. E.9), and a discussion of why we avoid hard-negative mining (App. B.2).

5 Conclusion

We proposed **GPS-CLIP**, a fine-tuning method that injects scene-graph structure through region- and relation-level *positive* objectives, without explicit hard-negative mining. By grounding alignment in regions, spans, and relations, GPS-CLIP constrains the visual manifold, mitigates modality asymmetry from caption-side hard negatives, and improves compositional reasoning while preserving zero-shot transfer and fine-grained understanding. UT-Zappos analysis further indicates improved object-attribute separability in the embedding space.

Limitations. GPS-CLIP is evaluated as a post-pretraining alignment stage, we make no claims about full web-scale pretraining. While the method is compatible with automatically generated structured supervision, its performance depends on the quality and diversity of boxes, region descriptions, and relation annotations. Our noise-robustness results suggest tolerance to moderately imperfect annotations, but larger and more diverse structured corpora may further improve spatial and relational reasoning. Broader scaling paths are discussed in App. F, and qualitative near-miss cases in App. G.2. Future work will explore GPS-CLIP alignment for diffusion models and other generative frameworks.

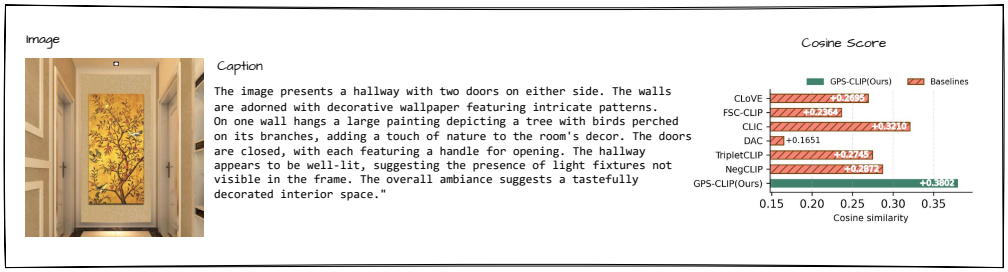


Figure 4: **Dense caption alignment.** Absolute image–caption cosine similarity for GPS-CLIP and baselines; GPS-CLIP scores highest. More examples in Appendix G.

Acknowledgments and Disclosure of Funding

Harsh Udai acknowledges support from the C-DOT STAR Scholarship Program during this research.

Competing interests. Vineeth N. Balasubramanian is also affiliated with Microsoft Research India. The authors declare no other competing interests.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL https://arxiv.org/abs/2102.05918.

[3] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[4] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zq1iJkNk3uN.

[5] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19265–19274, 2023. doi: 10.1109/CVPR52729.2023.01846.

[6] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

[7] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[8] Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation, 2025. URL https://arxiv.org/abs/2502.19842.

[9] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023. URL https://arxiv.org/abs/2310.19785.

[10] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

[11] Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining, 2023. URL https://arxiv.org/abs/2311.03964.

[12] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (dac) promote compositional reasoning in vl models, 2023. URL https://arxiv.org/abs/2305.19595.

[13] Yeyuan Wang, Dehong Gao, Lei Yi, Linbo Jin, Jinxia Zhang, Libin Yang, and Xiaoyan Cai. Enhancing fine-grained vision-language pretraining with negative augmented samples, 2024. URL https://arxiv.org/abs/2412.10029.

[14] Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in clip with efficient fine-tuning, 2025. URL https://arxiv.org/abs/2505.24424.

[15] Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. Bivlc: Extending vision-language compositionality evaluation with text-to-image retrieval, 2024. URL https://arxiv.org/abs/2406.09952.

[16] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations, 2023. URL https://arxiv.org/abs/2305.06152.

[17] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs, 2023. URL https://arxiv.org/abs/2305.06343.

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

[19] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021. URL https://arxiv.org/abs/2111.09734.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

[21] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

[22] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL https://arxiv.org/abs/2204.03162.

[23] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally?, 2023. URL https://arxiv.org/abs/2212.07796.

[24] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A. Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval, 2023. URL https://arxiv.org/abs/2305.03689.

[25] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. URL https://arxiv.org/abs/2306.14610.

[26] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little, 2021. URL https://arxiv.org/abs/2104.06644.

[27] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022. URL https://arxiv.org/abs/2205.14459.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

[29] Philipp J. Rösch, Norbert Oswald, Michaela Geierhos, and Jindřich Libovický. Enhancing conceptual understanding in multimodal contrastive learning through hard negative samples, 2024. URL https://arxiv.org/abs/2403.02875.

[30] Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. The hard positive truth about vision-language compositionality, 2024. URL https://arxiv.org/abs/2409.17958.

[31] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation, 2025. URL https://arxiv.org/abs/2501.09425.

[32] Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. Clip behaves like a bag-of-words model cross-modally but not uni-modally, 2025. URL https://arxiv.org/abs/2502.03566.

[33] Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained vlms for improving vision-linguistic compositionality, 2024. URL https://arxiv.org/abs/2410.05210.

[34] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding, 2024. URL https://arxiv.org/abs/2306.08832.

[35] Santiago Castro, Amir Ziai, Avneesh Saluja, Zhuoning Yuan, and Rada Mihalcea. Clove: Encoding compositional language in contrastive vision-language models, 2024. URL https://arxiv.org/abs/2402.15021.

[36] Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality, 2023. URL https://arxiv.org/abs/2305.13812.

[37] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining, 2021. URL https://arxiv.org/abs/2112.09106.

[38] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022. URL https://arxiv.org/abs/2112.03857.

[39] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding, 2022. URL https://arxiv.org/abs/2206.05836.

[40] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023. URL https://arxiv.org/abs/2311.06242.

[41] Yu-Guan Hsieh, Cheng-Yu Hsieh, Shih-Ying Yeh, Louis Béthune, Hadi Pour Ansari, Pavan Kumar Anasosalu Vasu, Chun-Liang Li, Ranjay Krishna, Oncel Tuzel, and Marco Cuturi. Graph-based captioning: Enhancing visual descriptions by interconnecting region captions, 2025. URL https://arxiv.org/abs/2407.06723.

[42] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

[43] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2818–2829. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/CVPR52729.2023.00276.

[44] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives, 2024. URL https://arxiv.org/abs/2411.02545.

[45] Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uAFHCZRmXk.

[46] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

[47] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts, 2022. URL https://arxiv.org/abs/2111.08276.

[48] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023. URL https://arxiv.org/abs/2207.00221.

[49] Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining, 2025. URL https://arxiv.org/abs/2502.14113.

[50] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training, 2021. URL https://arxiv.org/abs/2111.07783.

[51] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining, 2022. URL https://arxiv.org/abs/2204.14095.

[52] Haoxin Li and Boyang Li. Enhancing vision-language compositional understanding with multimodal synthetic data, 2025. URL https://arxiv.org/abs/2503.01167.

[53] Zexi Jia, Chuanwei Huang, Hongyan Fei, Yeshuang Zhu, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Jinchao Zhang, and Jie Zhou. A visual leap in clip compositionality reasoning through generation of counterfactual sets, 2025. URL https://arxiv.org/abs/2507.04699.

[54] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

[55] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

[56] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

[57] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

[58] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL https://arxiv.org/abs/1902.10811.

[59] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations, 2024. URL https://arxiv.org/abs/2406.11171.

[60] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 5571–5580. IEEE, October 2017. doi: 10.1109/iccv.2017.594. URL http://dx.doi.org/10.1109/ICCV.2017.594.

[61] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the proposed method and scope: GPS-CLIP adds grounded positive region-span and relation-triplet objectives during fine-tuning while preserving the standard CLIP dual-encoder inference pipeline. The method is detailed in Sec. 3, especially the region-level objective in Sec. 3.5, the triplet-level objective in Sec. 3.6, and the final structured fine-tuning objective in Sec. 3.7. The empirical claims are supported by the main benchmark results in Table 1 and Sec. 4, ablations in Sec. 4.3 and App. E, robustness/practicality analysis in Sec. 4.6, and limitations in Sec. 5. Compatibility beyond standard CLIP/InfoNCE is supported by the SigLIP-style results in Table 5 and App. E.8.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Sec. 5, noting that GPS-CLIP is evaluated as a post-pretraining alignment stage rather than during full web-scale pretraining, and that its performance depends on the quality of automatically generated scene-graph annotations such as boxes, region descriptions, and relation triplets. The paper also discusses scalability and practical deployment considerations in Sec. 4.6 and App. F, including inference-preserving behavior, training-time overhead, compatibility with silver-standard structured supervision, and robustness to moderately noisy annotations in Table 4 and App. E.6. App. G.2 further includes qualitative near-miss cases that illustrate remaining challenges in cluttered multi-actor scenes and highly detailed dense-caption settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results, theorems, or proofs. It introduces structured contrastive objectives and provides empirical validation, ablations, and representation analysis, the mathematical formulations of the losses are given in Sec. 3.1 and 3.4- 3.7.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the training setup in Sec. 4.1 and App. C, including the pretrained OpenCLIP backbone, GBC-1M split, filtering criteria, LoRA fine-tuning setup, optimizer, learning rate, batch size, number of epochs, warm-up, precision, distributed training configuration, and GPU type. The structured objectives are defined in Secs. 3.5- 3.7, the triplet extraction prompt is provided in App. C.3, and evaluation benchmarks and metrics are reported in Sec. 4 and App. D. Ablations and diagnostic settings are further detailed in App. E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Training and evaluation code, together with instructions for reproducing the main results, is publicly released at `https://github.com/Harsh-Udai/GPS-CLIP`. The paper further provides the full method description in Sec. 3, experimental setup in Sec. 4.1, implementation details in App. C, the triplet extraction prompt in App. C.3, and ablation/evaluation details in App. E.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Sec. 4.1 describes the dataset, backbone, baselines, benchmarks, and evaluation metrics, while App. C provides the training hyperparameters, optimizer, LoRA configuration, compute setup, preprocessing, and triplet extraction prompt. Detailed benchmark results and ablation settings are reported in App. D and E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper reports single-run results. This follows the common reporting convention in prior work on the same compositional vision-language benchmarks, where full multi-seed fine-tuning is computationally expensive. To improve reproducibility, we fix all training random seeds to 24, and evaluation is deterministic under the reported protocols. We support the claims through consistent trends across multiple benchmarks, controlled ablations, robustness analysis, and backbone/loss generalization experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: App. C specifies the training compute setup, including fp16 DDP training with HuggingFace Accelerate on 4 NVIDIA Tesla V100 GPUs, each with 32GB memory, batch size, and effective batch size. App. C.1 reports the one-time triplet extraction cost and discusses training-time overhead and unchanged inference-time latency, and FLOPs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The work uses existing vision-language datasets, pretrained models, and evaluation benchmarks for methodological and empirical analysis, and does not involve human-subject experiments, private user data, or deployment in high-stakes decision-making settings. We also discuss limitations, scalability considerations, and broader impacts in Sec. 5, App. F and App. H.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses broader impact in App. H. GPS-CLIP is a fine-tuning framework for improving compositional reasoning in CLIP-style vision-language models. Potential positive impacts include more reliable fine-grained visual understanding, better object-attribute binding, improved relation-aware retrieval, and reduced compositional mismatches in recognition or retrieval systems. The work does not generate open-ended content, introduce a deployed system, or directly interface with end users. Potential negative impacts are indirect and depend on downstream use, such as stronger vision-language retrieval or recognition systems being applied in sensitive content-analysis or surveillance settings.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new scraped dataset, high-risk generative model, pretrained language model, or deployed system. The work is methodological and uses existing datasets, pretrained models, and evaluation benchmarks for research on compositional vision-language alignment.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original creators of the datasets, benchmarks, pretrained models, and codebases used in the experiments, including OpenCLIP, GBC, Qwen2.5-7B-Instruct, HuggingFace Accelerate, and all evaluation benchmarks. We summarize the source and license/terms information for these assets in App. C.6, Table 9. We use all assets for academic research purposes and follow their original release terms, including non-commercial or research-use restrictions where applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our code is publicly released at `https://github.com/Harsh-Udai/GPS-CLIP` for reproducing the main training and evaluation results. The code is documented with instructions for environment setup, data preparation, training, and evaluation. The paper also documents the model, data, and preprocessing pipeline in App. C.6– C.7.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted using existing publicly available datasets, pretrained models, and evaluation benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects. All experiments are conducted using existing publicly available datasets, pretrained models, and evaluation benchmarks, therefore IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper uses Qwen2.5-7B-Instruct as a structured information extraction tool to convert existing captions into subject-relation-object triplets used for the relation-aware training objective. This use is part of the data preprocessing/training pipeline, not for generating scientific claims or evaluation labels. The model choice, motivation, prompt, and examples are documented in Appendix C.3, and the training setup using extracted triplets is described in Sec 4.1.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix Roadmap

A. Extended Related Work and Positioning

- **A.1.** Contrastive Pretraining and Compositional Failures
- **A.2.** Hard-Negative Methods and Their Trade-offs
- **A.3.** Structured Supervision and Scene-Graph Methods
- **A.4.** Region-Grounded and Architecture-Heavy Methods
- **A.5.** Supervision Distribution and Positioning of GPS-CLIP

B. Design Rationale and Geometry

- **B.1.** Positive-Centric vs Negative-Centric Training
- **B.2.** Why GPS-CLIP Avoids Hard-Negative-Driven Training
- **B.3.** Connection to CLIP Geometry
- **B.4.** Limitations of Prior Structured Approaches
- **B.5.** Design Space Summary

C. Implementation Details

- **C.1.** Training Setup and Computational Overhead
- **C.2.** Exact Implementation of the Region Objective
- **C.3.** Prompting Details for Triplet Extraction
- **C.4.** Relation Frequency vs. What'sUp Benchmark Relations
- **C.5.** Relation Diversity in Extracted Triplets
- **C.6.** Licenses and Existing Assets
- **C.7.** Model and Data Documentation

D. Extended Results

- **D.1.** What'sUp
- **D.2.** Winoground
- **D.3.** SugarCrepe
- **D.4.** SugarCrepe++
- **D.5.** COLA Multi-Object
- **D.6.** Zero-shot Classification and Retrieval
- **D.7.** MMVP-VLM

E. Ablations and Diagnostics

- **E.1.** Sensitivity to Loss Weights
- **E.2.** Diagnostic Freezing Ablation
- **E.3.** Effect of Weight Interpolation
- **E.4.** Relative Modality Gap, Variance, and PCA Geometry
- **E.5.** Causal and Probing Analyses
- **E.6.** Robustness to Noisy Scene-Graph Supervision
- **E.7.** Backbone Scaling to ViT-L/14
- **E.8.** SigLIP Compatibility and Design Discussion
- **E.9.** Region-Loss Component Ablation
- **E.10.** Directional Relation Diagnostics
- **E.11.** Controlled Baselines on the GBC Subset

F. Discussion on Scalability to Large-Scale Datasets

G. Qualitative Examples and Failure Modes

H. Broader Impact

A Extended Related Work and Positioning

A.1 Contrastive Pretraining and Compositional Failures

Contrastive pretraining has become the dominant paradigm for vision-language alignment, with models such as CLIP [1] and ALIGN [2] training dual encoders on hundreds of millions of web-scraped image-text pairs to maximize similarity between paired inputs while minimizing similarity with negatives. This simple yet scalable objective has produced universal cross-modal representations with strong zero-shot transfer across tasks, including image classification, retrieval, captioning, and even as backbones for large multimodal and generative models [19–21]. The effectiveness of this paradigm is closely tied to scale, massive datasets such as LAION-400M/5B [18] compensate for noise, enabling models to learn robust global alignments.

However, several studies reveal that CLIP-style contrastive objectives often encourage shortcut learning, models successfully separate easy negatives but fail on fine-grained distinctions involving attributes, relations, or word order, leading to "bag-of-objects" behavior rather than structured compositional understanding [22, 10]. Benchmarks such as Winoground [22], ARO [10], CREPE [23], What'sUp [9], COLA [24], and SugarCrepe [25] repeatedly show near-chance behavior on tasks requiring systematic generalization, spatial reasoning, or fine-grained linguistic distinctions. These failures indicate that global image-text alignment alone often ignores structured bindings between objects, attributes, and relations.

Extensions have attempted to address these weaknesses through data filtering or augmentation [13, 11], synthetic captions [44, 12], auxiliary objectives such as masked modeling and image–text matching [26, 28], and geometric regularization such as CyCLIP [27]. These methods improve robustness in different ways, but the core limitation remains: contrastive functions defined only over global embeddings lack the inductive bias required for compositional reasoning. GPS-CLIP addresses this by injecting structured region- and relation-level supervision directly into the contrastive framework.

A.2 Hard-Negative Methods and Their Trade-offs

Hard negatives-minimally perturbed captions or images that alter attributes, relations, objects, or word order-are widely used to improve compositionality [10, 13, 29]. They can be generated via rule-based swaps, scene graphs [16, 17], LLMs [11], or image-space generative editing. These strategies often raise fine-grained benchmark scores by forcing the model to distinguish near-miss captions or counterfactual images.

Loss engineering for hard negatives. FSC-CLIP [33] observes that naively adding text-side hard negatives can destabilize alignment: semantically close negatives may encourage global geometric drift rather than learning the intended distinction. It addresses this by defining token-patch-based local image-text matching and calibrating hard-negative supervision to account for ambiguity. CE-CLIP [34] constructs diverse hard negatives targeting multiple compositional factors, including attributes, objects, actions, and relations, and complements cross-modal contrast with intra-modal text separation and adaptive ranking-style constraints. Overall, these methods inject compositional structure through objective design, but their effectiveness depends strongly on the quality and faithfulness of synthetic hard-negative construction.

Image-space counterexamples. GNM-CLIP [11] generates hard negatives by minimally editing localized image regions, for example via diffusion-based inpainting, to change a targeted concept while preserving nuisance factors such as background, pose, and style. These edited samples act as semantically close impostors during contrastive training, increasing pressure on the representation to become sensitive to the edited factor instead of exploiting spurious correlations. This direction provides strong counterexamples, but introduces practical complexity through localization and generative editing, and can be sensitive to editing artifacts or distribution shift.

Despite their usefulness, hard negatives can introduce a compositionality-generality trade-off. Applying hard-negative objectives at the global embedding level can destabilize alignment, make models brittle to meaning-preserving edits, and hurt zero-shot recognition or retrieval [30]. Prior work also studies related failures involving negation and unimodal composition [31, 32]. In contrast, GPS-CLIP

does not rely on explicit hard-negative mining. It uses standard in-batch negatives as in vanilla CLIP, while scene-graph annotations define additional positive region-span and relation-level alignment constraints.

A.3 Structured Supervision and Scene-Graph Methods

Structured representations such as scene graphs provide a natural way to model objects, attributes, and relations, and have been applied to captioning, retrieval, and image generation [41]. In vision-language modeling, contrastive objectives remain highly effective [47, 9, 48], but standard CLIP-style training relies on global image-text embeddings, which tends to discard fine-grained structure and encourage "bag-of-objects" behavior.

Structured textual supervision via scene graphs. Structure-CLIP [16] uses scene-graph supervision to enrich textual representations and generate structured hard negatives. MosaiCLIP [36] converts captions into scene graphs and enumerates subgraphs at multiple granularities. These subgraphs are verbalized into multiple textual positives for the same image, yielding a multi-positive contrastive signal that encourages image embeddings to support multiple valid semantic decompositions rather than a single surface caption. MosaiCLIP also constructs structured hard negatives through graph perturbations such as relation swaps or attribute edits. While effective, this line remains primarily text-structure-driven and depends on the quality of scene-graph parsing and template-based verbalization.

Scene-graph modules and structured inference. SGVL [17] incorporates structure via architectural modifications, such as scene-graph prediction heads or graph tokens, and also relies on graph-based hard negatives. Such designs introduce stronger structured inductive biases, but may require additional components and structured inputs or predictions beyond the standard CLIP dual-encoder interface.

GPS-CLIP uses scene-graph supervision differently. We keep the CLIP dual encoder unchanged and use GBC-derived region-span pairs and $\langle s, r, o \rangle$ triplets solely to define positive region- and relation-aware contrastive losses in the shared embedding space. Scene-graph machinery is used only during training, and inference requires only raw images and text. Thus, GPS-CLIP is a drop-in, structure-enhanced alignment stage for CLIP-style models rather than an architecture-level scene-graph model.

A.4 Region-Grounded and Architecture-Heavy Methods

A related line of work introduces explicit region-level or object-centric structure into vision-language models. Region-grounding models, such as RegionCLIP [37], GLIP/GLIPv2 [38, 39], and Florence-2 [40], introduce region tokens, proposal mechanisms, or pixel-level supervision for fine-grained localization. Object-centric approaches such as OC-CLIP [49] target binding failures by constructing object-centric representations that competitively assign patches to object slots, discouraging many-to-many entanglement. OC-CLIP further enforces relation sensitivity by discriminating correct relational structure against perturbed variants, such as swapped subject/object roles, through a local relational objective.

These approaches directly address localization, binding, and relation modeling, but they often introduce additional architectural components, object-slot mechanisms, graph heads, or inference-time localization requirements. GPS-CLIP targets a complementary regime: rather than producing explicit localized outputs, it improves global image-text alignment by enforcing consistency between structured components during training. As a result, the model retains the standard dual-encoder interface and incurs no additional inference-time cost.

Prior formulations commonly resolve local matching through a selected region-concept correspondence, token-level maximum similarity, or alignment across multiple representation granularities. For example, RegionCLIP [37] selects one best concept per region, FILIP [50] uses token-wise maximum similarity without boxes, and PyramidCLIP [51] aligns across granularity levels while softening negatives for dataset-level noise; none of these assigns graded positives over overlapping same-image regions. In contrast, GPS-CLIP defines an overlap-aware assignment in which the designated region-span pair remains positive, compatible same-image overlaps are added as IoU-weighted positives,

other same-image pairs are masked, and cross-image pairs remain negatives. The contribution is thus the joint positive-assignment rule combining masking, multi-positive promotion, and IoU-calibrated weighting rather than region-level contrastive alignment in isolation.

A.5 Supervision Distribution and Positioning of GPS-CLIP

Several methods improve compositionality by modifying the supervision distribution. DAC [12] argues that compositional weakness is partly data-driven and densifies supervision by generating higher-quality, more detailed captions per image. It trains with a multi-instance cross-modal contrast that reduces reliance on any single caption. CLIC [14] manufactures compositional signal through combinatorial augmentation: concatenating two images and composing their captions to form multiple valid positives, while also creating structured hard negatives via controlled word or syntactic swaps. These strategies are effective in practice, though they either remain text-driven, where caption generation quality limits signal, or introduce an artificial input distribution through concatenated scenes.

A recurring empirical observation is that aggressively optimizing for compositional benchmarks can degrade general retrieval or recognition. CLoVe [35] addresses this through model patching, interpolating between pretrained and compositional fine-tuned checkpoints to recover generality while retaining some compositional gains. This provides a practical knob for trading off capabilities, but it does not directly specify how to ground compositional factors in the visual signal.

Synthetic counterfactual supervision. Recent methods also improve compositionality by generating synthetic counterfactual image-text data. SPARCL [52] generates LLM-edited positive/negative captions and T2I-generated images, using image feature injection and an adaptive margin loss to handle synthetic positives, hard negatives, easy negatives, and noisy generations. Visual Leap / GCS [53] constructs counterfactual sets by using LLMs to parse entities and relations, then applying block-based diffusion to add, remove, or modify visual elements as compositional "puzzle pieces," followed by a set-based training loss. These approaches are complementary to GPS-CLIP: they improve compositionality by synthesizing new counterfactual examples, whereas GPS-CLIP uses existing grounded scene-graph annotations to define positive region–span and relation-level alignment constraints over real images, avoiding the quality-control burden of synthetic image generation.

GPS-CLIP differs from these directions in where it injects compositional structure. Rather than relying primarily on templated subgraph verbalization [36], expensive generative image edits [11], caption densification [12], compositional image concatenation [14], or additional object-slot binding modules [49], GPS-CLIP leverages grounded region-level supervision available in Graph-Based Captions (GBC). We retain the standard global image–text contrastive objective while aligning region representations to matched textual spans/entities via a multi-positive grounded formulation with IoU-based weighting, and enforcing scene-level consistency with ordered subject-predicate-object triplet configurations. This grounding-centric design aims to address binding and relational errors at their source by linking language factors to localized visual evidence, while minimizing reliance on synthetic text templates, heavy caption synthesis pipelines, or post-hoc checkpoint patching.

Together, these comparisons position GPS-CLIP as a positive-centric, image-side structured alignment method that differs from hard-negative, text-augmentation, caption-densification, synthetic counterfactual generation, and architecture-heavy approaches while preserving the standard CLIP inference interface.

B Design Rationale and Geometry

B.1 Positive-Centric vs. Negative-Centric Training

Contrastive learning objectives can be viewed as balancing two forces: *repulsion* between negatives and *attraction* between positives. Most prior work incorporating structured supervision focuses on improving the repulsion term, for instance by constructing harder or more informative negatives through structured perturbations.

In contrast, GPS-CLIP modifies the attraction structure itself. Scene-graph topology is used to define multiple grounded positive constraints, including region-span correspondences and relation-level

consistency. Rather than only enforcing that matched image-text pairs are closer than mismatched pairs, the model is encouraged to satisfy a set of internally consistent alignment relationships.

This distinction has important geometric implications. Hard negatives primarily affect local decision boundaries by refining relative ordering between examples. Structured positives, on the other hand, directly shape the geometry of the embedding manifold by enforcing consistency across multiple views of the same underlying scene. This can lead to more stable and semantically coherent representations, particularly for compositional reasoning.

B.2 Why GPS-CLIP Avoids Hard-Negative-Driven Training

A central motivation for our formulation is the inherent *information imbalance* [45] between the two modalities in standard CLIP-style training. Captions explicitly encode objects, attributes, and relations, whereas images provide these structures only implicitly through spatial layout and appearance. A seemingly natural way to "resolve" this imbalance is to engineer increasingly descriptive captions and then generate families of hard negatives around them, for example via templates or large language models, so that the model learns to separate fine-grained textual variants. However, such strategies operate almost entirely on the text side of the joint space. Methods that rely heavily on hard textual negatives attempt to correct this imbalance by aggressively separating near-miss captions, such as swapped objects, attributes, or roles, but prior analyses show that this leads to an asymmetric embedding geometry, the text manifold becomes well-separated, yet the image manifold remains under-constrained, leaving visually fine-grained concepts entangled and producing unstable image-text retrieval behavior [15].

GPS-CLIP addresses this imbalance structurally. Rather than pushing images away from synthetically corrupted captions, we impose visually grounded constraints through *region-aligned multi-positives* and *relation-aware triplets*. These losses explicitly pull images toward their own part-level structure-subregions that satisfy IoU alignment and triplets that enforce role-specific consistency. This creates bidirectional geometric constraints that tighten the image manifold, disentangle object- and attribute-level variations, and yield a more stable and symmetric joint embedding space. In effect, GPS-CLIP strengthens the visual side of the representation, which hard-negative-based formulations typically leave weakly supervised.

This geometric improvement is corroborated by our analysis on **UT-Zappos-50K** in Sec. 4.4, a dataset known for subtle attribute distinctions such as shape vs. style and material vs. texture. On this benchmark, GPS-CLIP produces clearer separation between closely related attribute pairs, while hard-negative-driven variants exhibit weaker gains. This aligns with our central claim: improvements derived from linguistic perturbations do not reliably translate to visually grounded distinctions. By grounding supervision directly in visual evidence rather than synthetic caption edits, GPS-CLIP yields a more robust, balanced, and compositional embedding space across modalities. In this sense, GPS-CLIP complements hard-negative approaches by strengthening the attraction structure rather than further refining repulsion.

B.3 Connection to CLIP Geometry

Standard CLIP training enforces pairwise alignment between images and texts, encouraging matching pairs to be closer than mismatched ones. This can be viewed as a set of independent alignment constraints. GPS-CLIP extends this formulation by introducing *multi-view consistency constraints*. Each image-text pair is associated with multiple structured sub-components, such as regions, spans, and relations, and alignment is enforced not only at the global level but also across these components.

This can be interpreted as constraining the embedding space to satisfy a richer set of geometric relationships, potentially improving compositional generalization and robustness.

B.4 Limitations of Prior Structured Approaches

Existing approaches exhibit several common limitations.

Hard-negative reliance. Methods based on structured negatives depend on the quality of negative construction. Noisy or ambiguous negatives can introduce instability and increase gradient variance.

Text-centric structure. When structure is applied primarily on the text side, it assumes that captions fully capture the compositional structure of the scene, which may not hold in practice due to ambiguity or incompleteness.

Architectural overhead. Methods that incorporate structure via additional modules or tokens often incur increased model complexity and may require structured inputs at inference time.

GPS-CLIP mitigates these issues by shifting supervision toward structured positive constraints applied during training, without introducing additional architectural components or inference dependencies.

B.5 Design-Space Summary

Structured representations have been incorporated into vision-language models along several largely orthogonal design axes. We characterize prior work along three dimensions.

(i) Where structure is injected. Structure may be introduced on the *text side*, for example through caption augmentation or structured negatives; on the *image side*, for example through region tokens or pixel supervision; or implicitly in the *joint embedding space* through alignment constraints.

(ii) How structure is used. Existing methods primarily use structure to construct *hard negatives* or introduce *auxiliary prediction tasks*, such as graph prediction via additional heads. In contrast, structure can also be used to define *positive alignment constraints*, directly shaping the embedding geometry.

(iii) When structure is required. Some approaches rely on structure only during training, while others require structured inputs, such as region proposals or graphs, at inference time.

Table 6: **Positioning of structured supervision methods along key design axes.**

Method Type	Structure Location	Usage	Inference
Structure-CLIP / MosaiCLIP	Text	Negatives / aug.	No
SGVL-style methods	arch.	Aux. pred. + neg.	Partial
Region-grounding models	Image (regions)	Localization	Yes
GPS-CLIP (ours)	Joint embedding	**Positive constraints**	No

Within this space, GPS-CLIP occupies a distinct regime: structure is injected only through the training objective, used to define positive alignment, and not required at inference time.

C Implementation Details

We use a **ViT-B/32** model pretrained on **LAION-2B**, implemented with the **OpenCLIP** [43] library. The model is fine-tuned for **26 epochs** using **AdamW** (lr=1×10^{-5}, weight decay=0.01) with 500 warm-up steps followed by cosine decay.

Training is performed in **fp16** precision with distributed data parallel (DDP) using HuggingFace `accelerate` [54]. We use a batch size of **120 image-text pairs per GPU** across **4 GPUs**, corresponding to **480 global image-text pairs per optimization step**, with gradient accumulation set to **1**. The effective number of contrastive items for the structured losses is larger and sample-dependent, since each image-text pair contributes multiple region-span alignments and subject-relation-object triplets derived from its scene graph.

For efficient adaptation, we apply **LoRA adapters** (rank=16, $\alpha = 32$, dropout=0.1) to both text and vision transformers. Each training sample is decomposed into structured elements, with at least **5 region crops** paired with corresponding text spans and at least **10 subject-relation-object triplets** per image.

Our objective combines a global contrastive loss, a region-level IoU-masked contrastive loss, and a triplet contrastive loss. We emphasize that GPS-CLIP is a **fine-tuning stage** applied to a pretrained OpenCLIP model rather than training from scratch, and all baselines are trained under identical global image-text batch-size conditions for fair comparison.

Algorithm 1 IoU-weighted multi-positive region loss

Require: Region embeddings $\{\mathbf{v}_i\}_{i=1}^{N_r}$, span embeddings $\{\mathbf{t}_i\}_{i=1}^{N_r}$, boxes $\{b_i\}$, image ids $\{m_i\}$
Require: Temperature τ, IoU threshold τ_{pos}, designated-positive mass α, exponent γ
Ensure: Bidirectional region loss $\mathcal{L}_{\text{region}}$

$z_{ij} \leftarrow \text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau$ for all i, j
for $i = 1, \ldots, N_r$ **do** ▷ positive sets, weights, candidates
  $\mathcal{P}_i \leftarrow \{j \neq i : m_j = m_i \wedge \text{IoU}(b_i, b_j) \geq \tau_{\text{pos}}\}$
  if $\mathcal{P}_i \neq \emptyset$ **then**
    $w_{ij} \leftarrow (1-\alpha)\,\text{IoU}(b_i, b_j)^\gamma / \sum_{\ell \in \mathcal{P}_i} \text{IoU}(b_i, b_\ell)^\gamma$ for $j \in \mathcal{P}_i$; $\quad w_{ii} \leftarrow \alpha$
  else
    $w_{ii} \leftarrow 1$
  end if
  $\mathcal{C}_i \leftarrow \{i\} \cup \mathcal{P}_i \cup \{k : m_k \neq m_i\}$ ▷ mask other same-image pairs
end for
for $i = 1, \ldots, N_r$ **do**
  $\ell_i^{v \to t} \leftarrow -\log \dfrac{\sum_{j \in \{i\} \cup \mathcal{P}_i} w_{ij}\, e^{z_{ij}}}{\sum_{k \in \mathcal{C}_i} e^{z_{ik}}}$ ▷ region anchor, row i
  $\ell_i^{t \to v} \leftarrow -\log \dfrac{\sum_{j \in \{i\} \cup \mathcal{P}_i} w_{ij}\, e^{z_{ji}}}{\sum_{k \in \mathcal{C}_i} e^{z_{ki}}}$ ▷ span anchor, column i
end for
return $\mathcal{L}_{\text{region}} \leftarrow \left(\frac{1}{N_r}\sum_i \ell_i^{v \to t} + \frac{1}{N_r}\sum_i \ell_i^{t \to v}\right)$

All experiments are run on **NVIDIA Tesla V100 GPUs**, with 32GB memory each.

C.1 Training Setup and Computational Overhead

One-time Scene-graph Preprocessing. GPS-CLIP leverages existing region-level annotations in GBC, requiring no additional visual extraction. The only auxiliary step is LLM-based triplet extraction from captions, performed *once offline* with vLLM [55] in approximately 3 hours on 6×V100 GPUs. Once this extraction is completed, training proceeds as a standard contrastive pipeline, where scene-graph components are integrated solely through the loss objectives rather than architectural modifications.

Training and Inference Efficiency. GPS-CLIP maintains the original ViT-B/32 and Transformer text encoder architectures, utilizing parameter-efficient LoRA adapters. While processing multiple views (global images, region crops, and triplets) results in a ≈3× increase in wall-clock training time compared to vanilla CLIP fine-tuning, this cost is strictly confined to the training phase. **At inference, GPS-CLIP is identical to the original CLIP model**, requiring only a single global image-text forward pass. Consequently, it delivers improved compositional reasoning with zero additional latency, parameter growth, or FLOPs during deployment.

Comparison with Hard-Negative Baselines. The computational overhead of GPS-CLIP remains modest compared to hard-negative mining frameworks such as NegCLIP or TripletCLIP. These methods typically require additional pipelines to synthesize counterfactual negatives using LLMs or generative models, often introducing significant and recurring preprocessing costs.

In contrast, GPS-CLIP operates on *existing structured supervision* and does not require additional negative generation. In our setting, scene-graph annotations are obtained from the GBC pipeline, which itself scales automatically using MLLMs and open-vocabulary detectors, rather than manual labeling. Our method leverages these annotations directly to define grounded positive alignments.

Furthermore, GPS-CLIP is applied as a *post-pretraining fine-tuning stage*, using a filtered subset (400K samples) rather than requiring structured supervision at full pretraining scale. Combined with its robustness to annotation noise (Sec. 4.6), this makes the approach practical under realistic, large-scale, silver-standard supervision settings.

C.2 Exact Implementation of the Region Objective

This section specifies how IoU-weighted positives and same-image masking are implemented in the region-level objective (Eq. 3, Sec. 3.5).

Batch construction. For a minibatch of N image–caption pairs, we collect all designated region–span pairs from all images in the batch, $\{(\mathbf{v}_i, \mathbf{t}_i, b_i, m_i)\}_{i=1}^{N_r}$, where N_r varies across batches (each image contributes at least 5 pairs). Here b_i is the region's bounding box, m_i its source-image identifier, $\mathbf{t}_i$ the ℓ_2-normalized text embedding of the region's designated GBC description, and $\mathbf{v}_i$ the ℓ_2-normalized region embedding, obtained by encoding the region crop with f_θ ([CLS] token). Each region is paired with exactly one designated span, so regions and spans share the index i. The region–span logits are

$$z_{ij} = \mathrm{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau,$$

where τ is the learnable temperature.

Positive and candidate sets. For each anchor i, the compatible same-image overlaps are

$$\mathcal{P}_i = \{\, j \neq i : m_j = m_i,\ \mathrm{IoU}(b_i, b_j) \geq \tau_{\mathrm{pos}} \,\}.$$

$\mathcal{P}_i$ excludes the anchor itself; the positive set $\mathcal{P}(i)$ in Eq. 2 of the main text corresponds to $\{i\} \cup \mathcal{P}_i$. Because IoU is symmetric, $j \in \mathcal{P}_i \Leftrightarrow i \in \mathcal{P}_j$. With the cross-image candidates $\mathcal{X}_i = \{k : m_k \neq m_i\}$, the candidate set is

$$\mathcal{C}_i = \{i\} \cup \mathcal{P}_i \cup \mathcal{X}_i.$$

Same-image pairs outside $\{i\} \cup \mathcal{P}_i$ are therefore masked (excluded from the denominator) rather than treated as negatives.

Positive weights. These are the directional weights $w_{ij}^{r\to t}$ and $w_{ji}^{t\to r}$ in Eq. 3. When $\mathcal{P}_i \neq \emptyset$,

$$q_{ij} = \frac{\mathrm{IoU}(b_i, b_j)^\gamma}{\sum_{\ell \in \mathcal{P}_i} \mathrm{IoU}(b_i, b_\ell)^\gamma}, \qquad w_{ii} = \alpha, \qquad w_{ij} = (1-\alpha)\, q_{ij}, \quad j \in \mathcal{P}_i,$$

so that $w_{ii} + \sum_{j \in \mathcal{P}_i} w_{ij} = 1$. If $\mathcal{P}_i = \emptyset$, we set $w_{ii} = 1$ and the objective reduces to a single-positive InfoNCE term with same-image masking. Since the weights sum to one, the numerator never exceeds the denominator and each per-anchor loss is non-negative.

Directional losses. The region-to-text loss for anchor region i is

$$\ell_i^{v\to t} = -\log \frac{w_{ii}\, e^{z_{ii}} + \sum_{j \in \mathcal{P}_i} w_{ij}\, e^{z_{ij}}}{\sum_{k \in \mathcal{C}_i} e^{z_{ik}}}.$$

The text-to-region loss for anchor span j uses the transposed logits $z_{kj} = \mathrm{sim}(\mathbf{v}_k, \mathbf{t}_j)/\tau$, the same positive and candidate sets (indexed by the span's own region j), and weights computed with j as the anchor:

$$\ell_j^{t\to v} = -\log \frac{w_{jj}\, e^{z_{jj}} + \sum_{i \in \mathcal{P}_j} w_{ji}\, e^{z_{ij}}}{\sum_{k \in \mathcal{C}_j} e^{z_{kj}}}.$$

Note that $w_{ji} \neq w_{ij}$ in general, since each is normalized over its own anchor's overlap set.

Bidirectional sum. We sum the directions:

$$\mathcal{L}_{\mathrm{region}} = \left(\mathcal{L}_{\mathrm{reg}}^{v\to t} + \mathcal{L}_{\mathrm{reg}}^{t\to v}\right).$$

Eq. 3 (Sec. 3.5) in the main text shows the two directional terms, and the loss weights λ in Eq. 5 (Sec. 3.7) and App. E.1 are applied to losses with the same normalization.

Hyperparameters. We use $\tau_{\mathrm{pos}} = 0.5$, $\alpha = 0.5$, and $\gamma = 1$ in all main experiments. IoU is computed on axis-aligned bounding boxes. Thus the designated correspondence receives half of the positive mass, and the other half is distributed among accepted overlaps in proportion to their IoU.

C.3 Prompting Details for Triplet Extraction

We used a prompting strategy with Qwen2.5-7B-Instruct to extract subject-relation-object triplets from captions.

Motivation. Before finalizing the pipeline, we experimented with several prompting styles for dataset adaptation, including zero-shot extraction, chain-of-thought prompting, and different formatting constraints. Among these, the strategy described below consistently yielded the most accurate and structured triplets, balancing coverage with low noise.

Why Qwen2.5-7B-Instruct? We selected Qwen2.5-7B-Instruct for its balance of accuracy, efficiency, and open availability. Compared to larger LLMs, it provides strong instruction-following capability while being lightweight enough for multi-GPU parallelization across millions of captions. Qwen models have been optimized for multilingual and structured extraction tasks, making them well-suited for relation parsing without requiring additional fine-tuning. This allowed us to reliably obtain consistent triplets at scale with manageable computational cost.

The instruction given to the LLM was:

Prompting Strategy

```
You are an information extractor for long, detailed captions. From the
given text, extract atomic relational triplets in the EXACT format <subject,
relation, object>, one per line.

Rules:

1) Emit ONLY triplet lines (no numbering/explanations).

2) Split conjunctions into separate atomic triplets; keep each triplet
minimal.

3) Resolve obvious pronouns to antecedents when unambiguous.

4) Use verbs or short verb phrases for relations; prefer canonical forms.

5) Use has-attribute for colors/materials/states.

6) Spatial relations: on, in, under, above, below, behind, in front of, next
to, near, between, across, through, around, toward, beside, across from.

7) Structural relations: part-of, made-of, attached to, connected to,
supported by, covered by, surrounded by, enclosed by, filled with, contain(s).

8) Interaction/action: hold, wear, carry, ride, drive, eat, drink, play, look
at, talk to, chase, follow, sit on, stand on, hang from, open, close.

9) Temporal/weather: e.g., time , is , <value>; weather , is , <value>.

10) Quantity/comparison: has-quantity, taller than, bigger than, smaller
than.

11) Ownership/identity/purpose/causality: belong to, owned by, is-a, located
in, used for, caused by.

12) Max 20 triplets; prioritize salient, non-redundant relations.
```

Few-Shot Examples

```
Below we show representative input captions and the corresponding extracted
triplets used as few-shot demonstrations.

Caption: ‘‘A hiker wearing a yellow raincoat carries a backpack and holds a
walking stick.’’

Triplets:

<hiker , wear , yellow raincoat>
<hiker , carry , backpack>
<hiker , hold , walking stick>

Caption: ‘‘A storefront with bright neon signs has its glass door closed; the
interior lights are on and the floor is wet.’’

Triplets:

<signs , has-attribute , neon>
<storefront , has-attribute , bright>
<door , has-attribute , glass>
<door , has-attribute , closed>
<lights , has-attribute , on>
<floor , has-attribute , wet>

Caption: ‘‘A steel bridge with wooden planks is supported by concrete pillars;
cables are attached to the towers and connected to the deck, which is covered
by a thin layer of snow.’’

Triplets:

<bridge , made-of , steel>
<planks , made-of , wood>
<planks , part-of , bridge>
<bridge , supported by , pillars>
<pillars , made-of , concrete>
<cables , attached to , towers>
<cables , connected to , deck>
<deck , covered by , snow>
```

C.4 Relation Frequency vs. What'sUp Benchmark Relations

To better understand the gap in compositional reasoning performance, we compare the frequency of relations used in the What'sUp benchmark (focused on spatial relations) against our generated triplets. Table 7 shows that these relations occur far less frequently in our dataset than the dominant relations reported in Table 8.

Table 7: **Frequencies of What'sUp benchmark relations in our extracted triplets.** Counts are raw frequencies before canonicalization.

Relation	Count	Relation	Count
on	328,303	behind	11,184
in	116,906	next to	13,391
left	3,783	beneath	1,621
right	3,868	above	41,363
top	2,350		

We observe that while relations such as *has-attribute*, *wear*, or *include* dominate the extracted triplets (hundreds of thousands of occurrences), spatial relations central to What'sUp appear with much lower frequency (often by one to two orders of magnitude). This imbalance suggests that the training data would benefit from greater spatial-relation diversity, where distinguishing between fine-grained spatial configurations is critical.

C.5 Relation Diversity in Extracted Triplets

Our prompting pipeline produced a wide variety of relations. Table 8 reports the 50 most frequent relations (excluding *has-attribute*) out of the full set of 89,241 unique relation strings (before canonicalization). The long-tailed distribution indicates that while a small number of relations dominate, there is substantial coverage across diverse relation types.

Table 8: **Top-50 most frequent relations in extracted triplets (excluding *has-attribute*).** Counts are raw frequencies before canonicalization.

Relation	Count	Relation	Count	Relation	Count
suggest	439,826	part-of	50,749	on left	28,469
on	328,303	filled with	50,662	on right	27,626
wear	210,261	adorned with	46,380	create	27,467
include	178,263	dress in	44,154	position	26,936
is	175,521	stand on	44,115	for	26,926
made-of	169,732	possibly	42,785	appear to be	26,559
has	134,608	above	41,363	engage in	23,694
have	118,770	indicate	40,356	consist of	23,393
in	116,906	has-pattern	37,844	from	22,951
feature	116,290	add	34,686	appear	22,768
has-quantity	106,037	display	34,576	contrast with	20,942
convey	89,885	at	32,946	contain	20,839
hold	83,160	adorn	32,123	has-design	20,657
has-color	79,943	look at	32,028	reflect	20,499
with	76,013	focus on	31,623	provide	20,409

C.6 Licenses and Existing Assets

We use publicly available datasets, pretrained models, codebases, and evaluation benchmarks. Table 9 summarizes the primary existing assets used in this work, together with their source and license/terms information as reported by the original releases. We use all assets for academic research purposes and follow their original release terms. We use GBC-derived structured annotations for fine-tuning and follow the original GBC release terms, the underlying images remain governed by their original source licenses.

C.7 Model and Data Documentation

Model Card Summary.

- **Base model:** OpenCLIP ViT-B/32 pretrained on LAION-2B.
- **Adaptation:** LoRA fine-tuning on both vision and text transformers.
- **Training objective:** Global contrastive loss, IoU-weighted region-span loss, and relation-triplet loss.
- **Inference:** Unchanged CLIP-style dual encoder; no scene graphs, boxes, or triplets required.
- **Intended use:** Research on compositional vision-language alignment and retrieval.
- **Limitations:** Performance depends on the quality and diversity of automatically generated structured annotations; the method has not been evaluated in full web-scale pretraining.

Data Card Summary.

- **Training data:** GBC-1M-derived structured supervision.
- **Filtering:** Images with at least 5 subregions and at least 10 extracted triplets, yielding 400k training samples.
- **Annotations:** Region descriptions, bounding boxes, attributes, and relation triplets; triplets are extracted from captions using Qwen2.5-7B-Instruct.

Table 9: **Existing assets used in GPS-CLIP.** We summarize the main datasets, benchmarks, pretrained models, and codebases used in this work, along with their source and license/terms information as reported by the original releases.

Asset	Source / Creator	License / Terms
OpenCLIP	OpenCLIP [43]	MIT license.
LAION-2B pretrained OpenCLIP weights	OpenCLIP / LAION [43]	Follow original OpenCLIP model-card and release terms.
MS-COCO	COCO [56]	COCO annotations are CC BY 4.0; images follow original source terms.
Flickr30k	Flickr30k [57]	Apache 2.0 license, according to the public release used.
ImageNetV2	ImageNetV2 [58]	MIT license.
What'sUp	What'sUp [9]	MIT license.
COLA	COLA [24]	MIT license.
Winoground	Winoground [22]	Meta Images Research License; non-commercial research/evaluation use only.
SugarCrepe	SugarCrepe [25]	MIT license.
SugarCrepe++	SugarCrepe++ [59]	MIT license.
MMVP-VLM	MMVP-VLM [46]	MIT license.
UT-Zappos-50K	UT-Zappos-50K [60]	Public research dataset; follow original release terms.
Qwen2.5-7B-Instruct	Qwen team, Alibaba Cloud [61]	Apache 2.0 license.
HuggingFace Accelerate	HuggingFace [54]	Apache 2.0 license.
GBC 1M	Graph-Based Captioning dataset [41]	Annotations: CC BY-NC 4.0; CC12M source and original image rights apply.
ARO (VG-Relation)	ARO [10]	Code: MIT license. Built on Visual Genome annotations (CC BY 4.0).
SigLIP ViT-B/16 (WebLI) weights	Google [6]	Apache 2.0.

- **Evaluation data:** What'sUp, COLA, Winoground, SugarCrepe, SugarCrepe++, ImageNetV2, MS-COCO/Flickr30k retrieval, MMVP-VLM, and UT-Zappos.
- **Licenses:** Summarized in Table 9.

D Extended Results

This section discusses the scores on the sub-metrics across different compositional and fine-grained understanding benchmarks.

D.1 Scores on What'sUp benchmark

Table 10 reports comprehensive results on the What'sUp benchmark. We note that CLIP fine-tuned on COCO and GQA-Spatial achieves higher scores, largely because these datasets overlap with the evaluation distribution, giving that variant an advantage. By contrast, our model is not fine-tuned on either COCO or GQA, making the evaluation more fair but naturally yielding lower scores. Importantly, What'sUp explicitly forces the model to resolve relational perturbations, making relation diversity a key factor. Standard datasets rarely annotate relations in captions, which limits compositional coverage. Our triplet-based objective is well-suited for such cases, but its effectiveness depends on having diverse relational supervision, with richer relation diversity, relation-aware modeling can substantially improve performance on benchmarks like What'sUp.

D.2 Scores on Winoground benchmark

Winoground evaluates challenging compositional disambiguation through paired image–caption setups. Using the same CLIP-like backbone and evaluation protocol, we report accuracy on the text, image, and group subtasks, along with the overall average (Table 11). GPS-CLIP attains the highest average score (18.35%) among the CLIP variants considered, with improvements on both

Table 10: **What'sUp.** Spatial/relational reasoning with controlled caption flips (e.g., above/below). Scores are accuracies (%). Best/second-best per column are **bold**/underlined.

Model	What'sUp Subset A	What'sUp Subset B	COCO-spatial One-obj.	COCO-spatial Two-obj.	GQA-spatial One-obj.	GQA-spatial Two-obj.
CLIP ViT-B/32	30.3	31.6	43.7	51.1	46.5	47.4
FT on train COCO, GQA-spatial	28.2	25.2	**67.2**	**60.7**	**64.4**	**54.6**
FT on LAION-4M-prep	31.6	34.6	43.1	48.9	44.3	50.9
FT on LAION-4M-prep + neg. cap.	32.0	26.5	39.9	48.9	47.3	45.7
DAC-LLMs	28.40	34.56	42.10	51.14	48.71	50.52
DAC-SAM	27.18	27.21	47.66	54.09	46.81	51.20
CLoVE (w/o patching)	**34.71**	24.75	45.39	51.82	46.90	47.42
CE-CLIP	31.07	30.39	42.01	52.05	47.07	47.77
FSC-CLIP (CC3M)	29.85	29.41	42.01	51.36	46.81	46.39
CLIC-RedCaps	26.21	29.66	41.12	48.18	45.78	47.77
NegCLIP	32.5	36.3	47.4	46.4	45.3	46.7
TripletCLIP	26.9	32.6	46.5	51.5	56.65	32.0
GPS-CLIP (Ours)	34.46	**37.50**	50.24	55.23	55.66	51.19

Table 11: **Winoground results.** We report text, image, group scores (%) and their mean as **Avg.**. Higher is better. Best/second-best per column are **bold**/underlined.

Model	Text	Image	Group	Avg.
CLIP	30.70	10.50	8.00	16.40
DAC-LLMs	22.50	10.25	4.75	12.50
DAC-SAM	20.75	**12.25**	8.00	13.67
CLoVE (w/o patching)	22.00	9.50	6.50	12.67
CE-CLIP	19.25	12.00	5.25	12.17
FSC-CLIP (CC3M)	25.50	8.75	5.75	13.33
CLIC-RedCaps	31.25	12.00	**10.50**	17.92
NegCLIP	30.00	11.00	8.00	16.33
TripletCLIP	30.00	11.00	8.25	16.42
GPS-CLIP (Ours)	**34.80**	11.00	9.25	**18.35**

the text and group metrics. Notably, this gain is achieved without any form of hard-negative mining, indicating that the region- and triplet-aware supervision learned during fine-tuning also transfers to the fine-grained disambiguation required by Winoground.

D.3 Scores on SugarCrepe benchmark

Table 12 provides a comprehensive SugarCrepe comparison, covering both pretrained model and fine-tuned variants under the same Replace/Swap/Add perturbation protocol. This consolidated view enables direct, per-split inspection of how structured fine-tuning improves compositional robustness beyond global pretraining alone.

D.4 Scores on SugarCrepe++ benchmark

SugarCrepe++ extends SugarCrepe by pairing each image with two semantically correct but lexically diverse captions, along with a strong hard negative caption. The benchmark evaluates whether both positives consistently rank above the negative under several compositional perturbations (Table 13). Notably, SugarCrepe++ is designed to mirror natural human descriptions, where multiple paraphrastic interpretations of the same scene are equally valid. GPS-CLIP's strong performance therefore suggests that its structured supervision enables more human-aligned cross-modal reasoning, maintaining stable alignment across diverse yet semantically consistent captions.

Table 12: **SugarCrepe (SC) detailed results.** We report category-wise accuracies (%). Higher is better. Best/second-best per column are **bold**/underlined.

Model	Add Obj	Add Att	Replace Obj	Replace Att	Replace Rel	Swap Obj	Swap Att
CLIP	77.2	68.2	90.9	80.0	69.1	61.2	64.0
DAC-LLMs	89.6	**97.7**	94.4	89.2	**84.4**	**75.5**	74.2
DAC-SAM	87.6	95.2	91.2	85.8	83.9	71.8	74.6
CLoVE (w/o patching)	86.0	94.7	93.1	**91.2**	81.9	74.7	**77.6**
CE-CLIP	92.2	91.6	94.1	90.4	81.1	74.7	76.4
FSC-CLIP (CC3M)	87.1	93.2	92.4	85.8	77.0	70.6	71.6
CLIC-RedCaps	84.2	88.7	93.3	86.9	74.0	71.8	73.3
NegCLIP	86.95	88.44	91.53	83.25	73.97	72.24	67.72
TripletCLIP	90.40	86.27	94.43	85.53	80.94	69.80	69.82
GPS-CLIP (Ours)	**93.89**	95.66	**94.73**	85.15	72.11	71.00	70.50

Table 13: **SugarCrepe++ (SC++) results.** We report accuracies (%) for Replace (Obj/Att/Rel) and Swap (Obj/Att). Higher is better. Best/second-best per column are **bold**/underlined.

Model	Replace Obj	Replace Att	Replace Rel	Swap Obj	Swap Att
CLIP	87.89	68.02	50.14	37.95	48.19
DAC-LLMs	65.7	47.6	47.6	31.0	32.9
DAC-SAM	64.3	43.9	48.5	27.8	33.2
CLoVE (w/o patching)	73.7	63.3	55.1	35.5	45.2
CE-CLIP	81.8	62.4	53.0	40.0	40.2
FSC-CLIP (CC3M)	85.4	61.9	46.9	41.6	41.0
CLIC-RedCaps	90.1	**75.1**	**62.9**	**60.8**	**62.4**
NegCLIP	89.58	69.54	52.43	54.69	58.40
TripletCLIP	86.98	71.82	62.23	39.59	48.49
GPS-CLIP (Ours)	**92.88**	**75.1**	60.53	44.48	55.10

D.5 Detailed Cola Multi-Object Results

In the main paper, we report an *avg-only* summary of Cola to keep the presentation compact. Here we provide the full Cola Multi-Object breakdown, Table 14 reporting joint accuracies for T2I, I2T, and GROUP (AND). GPS-CLIP improves consistently across all three criteria, indicating stronger multi-object binding and reduced confusion under same-part distractors.

Table 14: **COLA multi-object.** We report T2I, I2T, and GROUP accuracies (%, higher is better). Best/second-best per column are **bold**/underlined.

Model	T2I	I2T	GROUP
CLIP	20.95	33.33	14.29
DAC-LLMs	24.29	19.52	11.43
DAC-SAM	16.67	21.90	8.10
CLoVE (w/o patching)	29.05	35.24	15.24
CE-CLIP	23.33	24.76	10.95
FSC-CLIP (CC3M)	16.19	27.14	9.05
CLIC-RedCaps	19.05	36.19	11.43
NegCLIP	16.19	32.38	11.43
TripletCLIP	10.95	32.38	9.05
GPS-CLIP (Ours)	**31.43**	**41.43**	**19.05**

D.6 Zero-shot Classification and Retrieval

Table 15 reports standard zero-shot transfer on ImageNetV2 (Top-1/Top-5) and bidirectional retrieval (Recall@1) on MS-COCO and Flickr30k. GPS-CLIP achieves the strongest overall performance,

improving retrieval Recall@1 across both datasets and directions while also attaining the best ImageNetV2 Top-1/Top-5 among all compared methods. These gains suggest improved global image-text alignment without sacrificing generalization to recognition-style evaluation.

Table 15: **Zero-shot classification and retrieval.** We report Recall@1 (%) on MS-COCO and Flickr30k for image-to-text (I2T) and text-to-image (T2I), and ImageNetV2 zero-shot Top-1/Top-5 (%). Higher is better. Best/second-best per column are **bold**/underlined.

Model	MS-COCO I2T R@1	MS-COCO T2I R@1	Flickr30k I2T R@1	Flickr30k T2I R@1	ImageNetV2 Top-1	ImageNetV2 Top-5
CLIP	50.02	30.35	78.59	59.72	55.96	83.41
DAC-LLMs	33.6	37.7	53.2	64.9	45.07	75.65
DAC-SAM	33.0	34.0	58.4	61.5	46.72	76.13
CLoVE (w/o patching)	48.0	42.5	69.2	68.2	46.05	75.10
CE-CLIP	55.3	**46.9**	74.9	68.3	43.32	74.28
FSC-CLIP (CC3M)	47.84	38.81	73.50	67.46	48.06	77.79
CLIC-RedCaps	52.5	33.8	79.4	62.8	55.30	82.49
NegCLIP	56.84	41.56	83.03	68.73	48.92	77.93
TripletCLIP	47.18	33.17	73.70	60.20	48.37	77.98
GPS-CLIP (Ours)	**57.54**	39.36	**85.50**	**69.27**	**58.11**	**83.67**

D.7 MMVP-VLM: Fine-grained Visual Pattern Understanding

Table 16: **MMVP-VLM:** Compositional accuracy (%) across visual patterns. Symbols denote visual attributes following the MMVP-VLM [46] taxonomy: Orientation and Direction, Presence of Specific Features, State and Condition, Quantity and Count, Positional and Relational Context, Color and Appearance, Structural Characteristics, Texts, and Viewpoint and Perspective. **Bold** indicates the best score and underline indicates the second-best.

Method	Backbone	Res										Avg.
CLIP	ViT-B/32	224	**26.7**	6.7	40.0	6.7	**26.7**	66.7	33.3	**26.7**	13.3	27.4
DAC-LLMs	ViT-B/32	224	13.3	20.0	26.7	**20.0**	6.7	53.3	**40.0**	**26.7**	6.7	23.7
DAC-SAM	ViT-B/32	224	0.0	20.0	26.7	6.7	13.3	66.7	26.7	**26.7**	20.0	23.0
CLoVE (w/o patching)	ViT-B/32	224	20.0	13.3	26.7	**20.0**	6.7	66.7	33.3	20.0	0.0	23.0
CE-CLIP	ViT-B/32	224	6.7	20.0	46.7	13.3	20.0	53.3	20.0	20.0	13.3	23.7
FSC-CLIP (CC3M)	ViT-B/32	224	6.7	6.7	13.3	**20.0**	6.7	53.3	26.7	**26.7**	20.0	20.0
CLIC-RedCaps	ViT-B/32	224	6.7	**26.7**	20.0	6.7	0.0	60.0	20.0	6.7	26.7	19.3
NegCLIP	ViT-B/32	224	20.0	13.3	46.7	13.3	20.0	46.7	20.0	**26.7**	26.7	25.9
TripletCLIP	ViT-B/32	224	6.7	13.3	40.0	6.7	13.3	53.3	13.3	13.3	20.0	20.0
GPS-CLIP (Ours)	ViT-B/32	224	**26.7**	20.0	**60.0**	6.7	20.0	**80.0**	**40.0**	13.3	**33.3**	**33.3**

We further evaluate compositional visual understanding on **MMVP-VLM** [46], which probes *fine-grained visual patterns* beyond standard retrieval/classification. MMVP-VLM groups questions by nine attribute categories: *Orientation and Direction*, *Presence of Specific Features*, *State and Condition*, *Quantity and Count*, *Positional and Relational Context*, *Color and Appearance*, *Structural Characteristics*, *Texts*, and *Viewpoint and Perspective*. For each method, we report accuracy (%) per category and their **macro-average** (**MMVP-VLM Avg**) across categories.

Table 16 shows that *GPS-CLIP* yields the strongest overall performance, improving the MMVP-VLM macro-average from 27.42 (CLIP) to **33.33**. Gains are particularly pronounced for *Color and Appearance* (66.7 → 80.0) and *Structural Characteristics* (33.3 → 40.0), while also improving *Viewpoint and Perspective* (13.3 → 33.3). These improvements suggest that region- and relation-aware supervision strengthens *localized compositional cues* and *attribute binding* that are typically under-specified by web-scale captions, complementing the global alignment trends observed on retrieval and classification benchmarks.

Under 20% joint scene-graph corruption (Table 4), the MMVP decline concentrates in categories most tied to structured supervision: Color and Appearance (12→9 correct pairs), Structural Characteristics

(6→3), and Positional and Relational Context (3→1), while the remaining six categories do not decline. Color and Structural Characteristics fall below pretrained CLIP (9 vs. 10 and 3 vs. 5), consistent with incorrect positive supervision being more harmful than absent supervision. Since MMVP-VLM contains only 135 pair-level decisions, each decision changes the average by 0.74 points, so a few binding errors produce a large-looking percentage drop.

E Ablations and Diagnostics

E.1 Sensitivity to Loss Weights

We vary the relative weighting $(\lambda_{\text{global}}, \lambda_{\text{region}}, \lambda_{\text{relation}})$ of the global, region-level, and triplet-level (relation) losses, respectively, to test robustness to hyperparameter choices. This ablation examines whether GPS-CLIP depends on a narrowly tuned balance between preserving pretrained global alignment, improving fine-grained region grounding, and enforcing relation-aware structure. As shown in Table 17, performance remains strong across several reasonable settings, with the best average result obtained when all three terms are weighted equally. Increasing the region weight improves the average over the first setting, suggesting that fine-grained benchmarks benefit from stronger local grounding. Increasing the triplet weight alone is also beneficial, but is less stable than the region-heavy setting, consistent with relation supervision being especially useful for relation-sensitive benchmarks such as What'sUp and COLA. In contrast, reducing or over-weighting the global loss hurts performance, indicating that structured supervision works best when it complements, rather than replaces or overwhelms, the pretrained CLIP alignment.

Table 17: **Sensitivity to loss weights.** Results showing average performance under different weight settings of the global (λ_{global}), region (λ_{region}), and relation ($\lambda_{\text{relation}}$) losses. Avg. is computed over What'sUp, COLA, Winoground, SugarCrepe, SugarCrepe++, and MMVP-VLM.

λ_{global}	λ_{region}	$\lambda_{\text{relation}}$	Avg. Acc. (%)	Notes
1.0	0.5	0.5	44.20	Reduced structured weights
1.0	1.0	0.5	45.80	Higher region weight
1.0	0.5	1.0	44.90	Higher triplet weight
0.5	1.0	1.0	43.60	Lower global weight
1.5	0.5	0.5	43.00	Overweighted global
1.0	1.0	1.0	**46.44**	All equal (default)

E.2 Diagnostic Freezing Ablation

We conduct a diagnostic ablation comparing two single-branch adaptation strategies during fine-tuning: (i) freezing the text encoder and training only the vision branch, and (ii) freezing the vision encoder and training only the text branch. All runs use the same structured supervision and training recipe, changing only which encoder branch is frozen. Table 18 reports results across retrieval and compositional probes.

We observe different behavior across task families. For retrieval tasks such as Flickr30k and COCO, freezing the text encoder and updating the vision branch yields stronger performance, suggesting that vision-side adaptation is useful for instance-level image–text matching. In contrast, for several compositional probes, especially SugarCrepe and parts of What'sUp/VGQA-style evaluation, freezing the vision encoder and updating the text branch performs better. This does not imply that the text branch alone is sufficient for compositionality; rather, under single-branch adaptation, language-side updates appear more beneficial for fine-grained caption discrimination involving attributes, relations, and role changes.

These trends are consistent with our broader representation analysis on UT-Zappos (Sec. 4.4), where fine-grained compositional performance is linked to the structure of both image and text embedding spaces. However, this frozen-encoder ablation is diagnostic and should not be interpreted as isolating a single causal branch for all compositional gains. Our best overall results are obtained when both encoders are fine-tuned with the full GPS-CLIP objective, allowing region- and relation-level supervision to shape the joint embedding space on both sides.

Table 18: **Diagnostic freezing ablation.** We compare two single-branch adaptation settings: training only the vision branch with the text encoder frozen, and training only the text branch with the vision encoder frozen. Retrieval performance is reported as Recall@1 (%) for Flickr30k and COCO; compositional performance is measured on What'sUp/VGQA-style probes and SugarCrepe. The full GPS-CLIP model fine-tunes both encoders.

Task / Benchmark	Text Frozen (training vision)	Vision Frozen (training text)
Flickr30k I2T	84.81	77.00
Flickr30k T2I	69.21	66.00
COCO I2T	59.02	48.50
COCO T2I	39.52	36.95
VGQA (one-obj)	46.66	56.40
VGQA (two-obj)	52.31	42.44
Controlled Set A	29.85	30.10
Controlled Set B	29.66	32.60
SugarCrepe (avg.)	77.32	81.13

Table 19: **Effect of weight interpolation on compositional benchmarks, ImageNetV2 classification, and zero-shot retrieval.** The non-interpolated model is the raw checkpoint obtained after structured fine-tuning, while the interpolated checkpoint is used for GPS-CLIP results. Retrieval entries report Recall@1.

	Compositional benchmarks						ImageNetV2		COCO		Flickr30K	
GPS-CLIP Checkpoint	What'sUp	COLA	Wino.	SC	SC++	MMVP	Top-1	Top-5	I2T	T2I	I2T	T2I
Raw fine-tuned	**47.60**	**33.97**	17.25	**84.07**	**66.10**	32.59	54.90	80.70	52.86	34.00	80.90	63.96
Interpolated	47.38	30.64	**18.35**	83.29	65.62	**33.33**	**58.11**	**83.67**	**57.54**	**39.36**	**85.50**	**69.27**

E.3 Effect of Weight Interpolation

To balance compositional performance and general-purpose zero-shot transfer, we interpolate the fine-tuned GPS-CLIP parameters with the pretrained parameters as $\theta = 0.25\,\theta_{\text{pre}} + 0.75\,\theta_{\text{ft}}$. Unless otherwise stated, the GPS-CLIP results reported use this interpolated checkpoint. To distinguish the effect of structured fine-tuning from that of interpolation, Table 19 also reports the raw fine-tuned checkpoint before interpolation.

The raw GPS-CLIP checkpoint attains a slightly higher average across the six compositional benchmarks than the interpolated checkpoint (Table 19) (46.93 versus 46.44), showing that the compositional improvements arise from structured fine-tuning rather than from interpolation. Interpolation instead provides a more balanced operating point; relative to the raw checkpoint, it improves ImageNetV2 Top-1 by 3.21 points and Top-5 by 2.97 points, while improving all four zero-shot retrieval directions. The average retrieval Recall@1 consequently increases from 57.93 to 62.92, at the cost of only 0.49 points in average compositional performance.

E.4 Relative Modality Gap, Variance, and PCA Geometry

To characterize how *integrated* the image and text embedding spaces are (beyond downstream accuracy), we analyze the geometry of similarities via the *Relative Modality Gap* (RMG). Concretely, we compute cosine-similarity statistics for: (i) **Match Cos** (paired image-text), (ii) **Img-Img Cos** (intra-image similarities), and (iii) **Txt-Txt Cos** (intra-text similarities). RMG summarizes the *distributional separation* between intra-modal cohesion and cross-modal alignment: a larger gap typically indicates that image and text embeddings form more *separate modality "islands"* (high within-modality similarity but weak cross-modal bridges), whereas a smaller gap suggests stronger neighborhood mixing across modalities. Table 20 reports these statistics.

What the numbers imply about embedding-space variation. The key failure mode we look for is not simply low Match Cos, but *imbalanced structure*: one (or both) modalities becoming overly

Table 20: Relative Modality Gap (RMG) and cosine similarity statistics across models.

Method	Match Cos	Img-Img Cos	Txt-Txt Cos	RMG
DAC-LLMs	0.2087	0.3370	0.4883	0.5740
FSC-CLIP (CC3M)	0.2913	0.4269	0.5701	0.5856
CLIC-RedCaps	0.2844	0.4679	0.5638	0.5965
NegCLIP	0.2847	0.5021	0.5092	0.5913
TripletCLIP	0.2598	0.6128	0.8444	0.7317
GPS-CLIP (Ours)	0.3690	0.4296	0.7703	0.6120

clustered/anisotropic while cross-modal matches remain comparatively weak. This imbalance makes compositional generalization harder because fine-grained edits (attribute swaps, relation flips) require the matched cross-modal neighborhood to be sensitive, not just the unimodal clusters to be tight.

I. **TripletCLIP** exhibits the **largest RMG** (0.7317): it produces very strong intra-modal cohesion (Img-Img: 0.6128, Txt-Txt: 0.8444) but relatively weak paired alignment (Match: 0.2598). Geometrically, this suggests two well-formed unimodal manifolds with insufficient cross-modal overlap-a modality-separation regime that can be brittle under compositional perturbations.

II. **NegCLIP** reduces RMG (0.5913) with a modest increase in Match Cos (0.2847), but it does so alongside **reduced intra-modal cohesion** (Img-Img: 0.5021, Txt-Txt: 0.5092). This pattern is consistent with a "gap reduction by loosening structure" effect: the modalities are less separated partly because unimodal neighborhoods are less consistent, which can hurt robustness for fine-grained binding.

III. **DAC / FSC-CLIP / CLIC** occupy a middle regime (RMG $\approx$ 0.57-0.60), with moderate Match Cos ($\approx$0.21-0.29) and moderate intra-modal cohesion. These methods improve compositional sensitivity relative to vanilla CLIP variants, but still show noticeable separation between matched pairs and unimodal similarity structure.

IV. **GPS-CLIP (Ours)** achieves the **highest Match Cos** (0.3690), indicating substantially stronger cross-modal bridges. At the same time, it preserves **strong text-space structure** (Txt-Txt: 0.7703) while keeping **image-space cohesion moderate** (Img-Img: 0.4296), avoiding overly tight image-only clustering. Although RMG is not the smallest (0.6120), the *shift* in Match Cos suggests improved cross-modal neighborhood overlap *without* collapsing unimodal geometry-a balance that is important for compositional edits where textual role/order must remain well-structured while visual evidence stays discriminative.

Takeaway. RMG should be interpreted together with the underlying similarity components: a small gap can arise either from *better cross-modal alignment* (desirable) or from *weakened intra-modal structure* (often undesirable). Across methods, stronger compositional generalization aligns with embedding spaces that (i) increase matched image-text similarity and neighborhood overlap, while (ii) maintaining coherent unimodal organization (especially on the text side for relation/role sensitivity). Our region-span and relation-aware objectives push the space toward this integrated-but-structured regime.

Table 21: **Intra variance in image and text embeddings (UT-Zappos-50k).** Higher intra-class variance suggests better distributed representations (less collapse). Note that the entries need to be multiplied by 10^{-5}. **Bold** = best, underline = second-best per column.

Method	Img-Obj	Img-Attr	Txt-Obj	Txt-Attr
DAC-LLMs	**46.90**	**40.66**	<u>7.23</u>	0.00
FSC-CLIP (CC3M)	34.62	29.71	6.61	0.00
CLIC-RedCaps	32.99	28.88	5.73	0.00
NegCLIP	31.27	26.95	6.86	0.00
TripletCLIP	23.07	19.45	4.43	0.00
GPS-CLIP (Ours)	<u>41.25</u>	<u>33.98</u>	**9.85**	0.00

Intra-class variance and embedding collapse (UT-Zappos). Table 21 reports *intra-class* embedding variance for object/attribute groups in the image and text spaces. Intuitively, higher variance indicates that instances within a semantic class occupy a *broader neighborhood* rather than collapsing to a tight cluster, which is desirable for fine-grained discrimination and robust compositional recombination. We observe that *GPS-CLIP* yields strong variance on the **text side** (best **Txt-Obj**), suggesting richer object semantics and reduced textual collapse compared to hard-negative and triplet-only variants. On the **image side**, *GPS-CLIP* is consistently second-best for both Img-Obj and Img-Attr, indicating improved distributional spread while avoiding excessive dispersion that can harm retrieval alignment. Notably, **Txt-Attr** variance is 0.00 for all methods, which is consistent with UT-Zappos prompts where attribute phrases are highly templated and thus induce near-identical attribute text embeddings; consequently, this metric primarily distinguishes methods through image-side attribute structure and text-side object structure.

Qualitative PCA visualization of hierarchical object-attribute geometry. To complement the cosine-similarity and variance analyses above, we provide a qualitative PCA visualization of the UT-Zappos embedding space. Figure 5 compares GPS-CLIP and NegCLIP using selected parent footwear categories, *Boots*, *Sandals*, *Shoes*, and *Slippers*, together with their associated fine-grained child concepts. Stars denote parent object embeddings, circles denote child concepts, and dashed lines indicate the assigned parent–child relationships.

The visualization suggests that GPS-CLIP yields a somewhat more coherent object-centered organization than NegCLIP. This is most visible for *Boots*: in GPS-CLIP, the parent anchor is more clearly separated from the other object anchors, and related concepts such as *Knee High*, *Over the Knee*, *Mid-Calf*, and *Ankle* occupy a relatively consistent region. This provides a qualitative sanity check for the quantitative trend reported in Sec. 4.4 and Table 3, where GPS-CLIP shows reduced object–attribute mixing while maintaining fine-grained alignment.

NegCLIP, in comparison, appears more dispersed in this projection. Several boot-related concepts remain distinguishable, but they are less consistently organized around the *Boots* anchor. This is consistent with the view that hard-negative-centric training may improve sensitivity to phrase-level differences without necessarily preserving a clean parent–child geometry.

Both models also show ambiguity among *Shoes*, *Sandals*, and *Slippers*, which is expected given their semantic and visual overlap. Some developmental footwear terms, such as *Firstwalker*, *Prewalker*, and *Crib Shoes*, also appear relatively far from their nominal parent categories, suggesting that the embedding space may encode factors beyond object identity, such as age group or usage context.

Since PCA projects high-dimensional embeddings into two dimensions, this figure should be interpreted as a qualitative diagnostic rather than a standalone measure of hierarchy preservation.

E.5 Causal and Probing Analysis

Table 22: **Effect of region- and triplet-level losses.** We report average accuracies (%) over compositional benchmarks (WhatsUp, COLA, Winoground, SugarCrepe, SugarCrepe++) and the fine-grained MMVP-VLM diagnostic, under different loss configurations. ✓ indicates that the corresponding loss is used. For each column, the best result is in **bold** and the second best is underlined.

Model	Region	Triplet	WhatsUp	COLA	Wino-ground	Sugar-Crepe	Sugar-Crepe++	MMVP-VLM	Avg.
CLIP			41.8	22.8	16.4	72.9	54.6	27.42	39.32
CLIP-GBC (global)			41.9	27.9	17.1	77.6	57.6	28.89	41.83
CLIP-GBC + Reg.	✓		44.6	30.6	17.8	80.7	63.27	28.89	44.31
CLIP-GBC + Tri.		✓	43.4	**33.0**	15.41	81.0	62.20	30.37	44.23
GPS-CLIP (Ours)	✓	✓	**47.38**	30.64	**18.35**	**83.29**	**65.62**	**33.33**	**46.44**

To disentangle the contribution of structure-aware learning from the effect of training on GBC data alone, we perform an ablation in which we first fine-tune a CLIP model using only global image-caption pairs from GBC (**CLIP-GBC (global)**), and then incrementally add the region-level and triplet-level losses. Table 22 reports results across all compositional benchmarks, including the MMVP-VLM diagnostic.

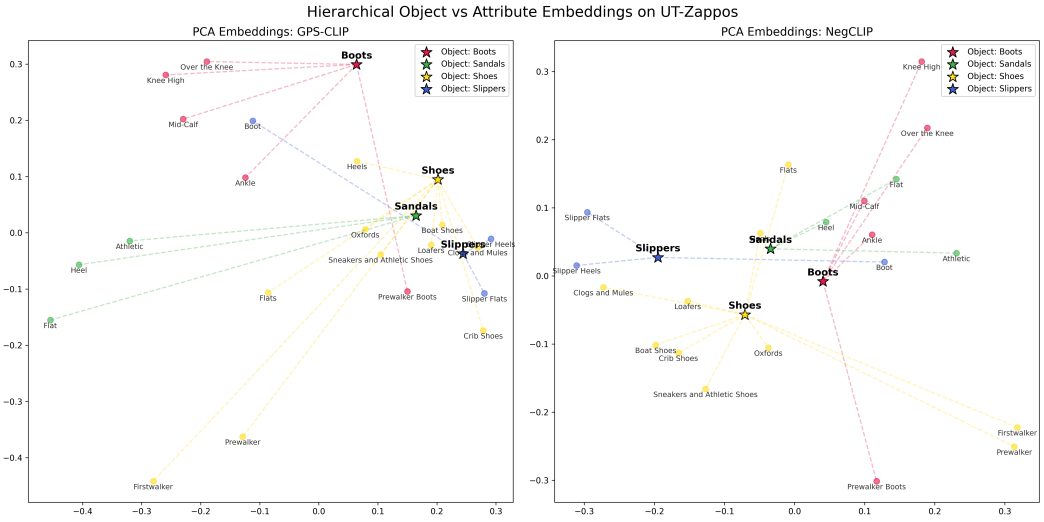


Figure 5: Qualitative PCA visualization of hierarchical object-attribute embeddings on UT-Zappos. We compare GPS-CLIP and NegCLIP using object-level anchors and associated fine-grained child concepts for *Boots*, *Sandals*, *Shoes*, and *Slippers*. Stars denote parent object embeddings, circles denote child concepts, and dashed lines connect each child concept to its assigned parent object. GPS-CLIP exhibits a more coherent object-centered organization, especially for *Boots*, whereas NegCLIP produces a more dispersed structure. Both methods show ambiguity among semantically close footwear categories such as *Shoes*, *Sandals*, and *Slippers*. Because the plot is based on a two-dimensional PCA projection, it should be interpreted as qualitative evidence.

Both structured objectives individually improve over the CLIP-GBC baseline: adding only the region-level loss increases average accuracy from 41.83% to 44.31%, and adding only the triplet-level loss yields a similar improvement to 44.23%. When combined, these two objectives produce the full GPS-CLIP model, which achieves the highest overall accuracy (46.44%). The gains consistently appear on benchmarks that require attribute binding, relation reasoning, and multi-positive alignment, indicating that the improvements stem directly from enforcing region- and relation-aware alignment rather than from data effects alone. This provides causal evidence that the structured losses are responsible for the additional compositional capacity observed in GPS-CLIP.

Ablation overview (three guiding questions). We structure our ablations around three questions, each targeting a different alternative explanation for the gains.

1. **Data vs. objective: Is this just better supervision?** A common concern is that improvements may come purely from training on more/better data, rather than from the proposed structured learning. To isolate this, we introduce **CLIP-GBC (global)**, which fine-tunes CLIP using *only* global image-caption pairs from GBC, without any region or relation objectives. This measures the *data/supervision* effect alone: CLIP→CLIP-GBC (global) yields a modest gain (39.32 → 41.83, +2.51 Avg.). Crucially, we then add structured objectives on top of the *same* CLIP-GBC baseline, holding the data fixed. Adding **Region-only** improves Avg. to 44.31 (+2.48 over CLIP-GBC), and adding **Triplet-only** improves Avg. to 44.23 (+2.40). When *both* are used, **GPS-CLIP** reaches 46.44 (+4.61 over CLIP-GBC). Because these gains are obtained without changing the underlying data (only the objectives), they provide direct evidence that improvements are *not* explained by better supervision alone, but arise from the proposed structure-aware constraints.

2. **Component vs. recombination** (linking Table 1 and Table 22). Table 22 provides a controlled decomposition of *which supervision signals matter*, while Table 1 places the same ingredients in the broader benchmark landscape. Importantly, although our Region and Triplet objectives are related to prior directions (local alignment and relational constraints), they are **not instantiated in the same way** in existing CLIP variants.

   Region objective (conceptually related to local alignment, but not prior token-patch losses). The Region-only variant in Table 22 injects compositional structure by aligning *grounded re-*

gion representations to matched textual spans/entities using **a multi-positive, IoU-weighted formulation** derived from GBC supervision. This differs from prior token-patch matching or hard-negative-aware local losses (e.g., FSC-CLIP/CE-CLIP), which primarily operate through token-level heuristics and/or synthetic hard negatives without explicit region-span grounding. Consistent with its grounding-centric role, Region-only improves over CLIP-GBC(global) (Avg. 41.83→44.31), with strong gains on structure-sensitive compositional benchmarks (e.g., WhatsUp and SugarCrepe/SC++).

Triplet objective (conceptually related to relation supervision, but not prior triplet consistency). Similarly, the Triplet-only variant aligns the full-image embedding with ordered subject–predicate–object triplets extracted from captions, providing a scene-level relational inductive bias rather than applying generic triplet constraints or purely text-driven relation perturbations. This yields a comparable improvement (Avg. 41.83→44.23), and strengthens relational discrimination as reflected by gains on compositional diagnostics and the role-swap probes in App. E.10.

GPS-CLIP is not a naive sum of losses. Our gains do not come from simply assembling previously standard losses; rather, they follow from injecting grounded structure as positives at multiple granularities. Table 22 shows that adding either grounded *region-span* or *relation-triplet* alignment already improves over the same CLIP-GBC(global) baseline, and combining them yields the strongest overall Avg. (**46.44** vs. <u>44.31</u> and 44.23), indicating that these signals are complementary. Table 1 further suggests that this **grounded, positive-centric encouragement** delivers a more *balanced* improvement than hard-negative-driven strategies: while HN-based objectives can amplify sensitivity through *synthetic counterexamples* or token-level heuristics, they often trade off generalization or overfit to perturbation templates. In contrast, GPS-CLIP encourages compositionality by grounded supervision to the visual evidence (regions and their relations), strengthening binding and role-awareness without relying on extensive hard-negative construction.

3. **Design-choice necessity:** Does our specific formulation matter? We ablate how structure is injected in the region objective, targeting (a) same-image masking, (b) multi-positive promotion of overlapping regions, and (c) IoU-based weighting. These are not cosmetic: prior structure-aware approaches often enrich the text side by parsing scene graphs and generating fine-grained attribute/relation edits to construct hard negatives (e.g., Structure-CLIP/MosaiCLIP), or introduce additional modules (e.g., slot-centric/object-token mechanisms) to explicitly represent entities. In contrast, GPS-CLIP follows a more perceptual decomposition: first anchor entities via region-level alignment, then learn inter-entity relations via explicit (s, r, o) consistency without auxiliary heads/tokens and without relying on synthetic hard-negative templates. As shown in Table 25 (App. E.9), each component yields an incremental gain (+1.46, +0.36, +2.74), and App. E.10 (Table 27) shows that the triplet objective induces role-order sensitivity, confirming that the specific inductive biases are consequential. Beyond these components, the representation analysis (App. B.2) indicates that grounded positive supervision tightens both the image and text manifolds, yielding a more balanced joint embedding space than caption-side hard-negative training.

E.6 Robustness to Noisy Scene-Graph Supervision

Scope. GPS-CLIP relies on usable grounded supervision during its post-pretraining alignment stage, we do not claim that it is independent of annotation quality or coverage. The GBC region descriptions and boxes are generated using open-vocabulary detectors and MLLMs, while relation triplets are extracted from captions using Qwen2.5-7B-Instruct. These annotations should therefore be regarded as automatically generated silver-standard supervision. Importantly, structured annotations are required only during fine-tuning, inference retains the unchanged dual-encoder architecture and requires no boxes, region descriptions, or scene graphs.

Noise protocol. Table 4 shows corruptions at 10% and 20% of the region and triplet supervision. We extend this analysis by corrupting the two annotation streams separately and by evaluating a severe 50% joint-corruption condition. Region corruption replaces the description of a selected region with that of a randomly sampled region. Triplet corruption applies one of the following to a selected triplet: replacing the relation with a randomly sampled relation, swapping the subject and object roles, or replacing an attribute value with a randomly sampled one. The corrupted k% of items are sampled uniformly at random. The data, backbone, global objective, optimizer, training schedule, and structured-loss weights are held fixed. These interventions therefore probe sensitivity to region-description and relation quality rather than detector localization error.

Table 23: **Controlled robustness tests for noisy structured supervision.** Region and triplet corruption are varied independently while all other training settings remain fixed. The CLIP-GBC global model does not use the corrupted structured annotations and is included as a reference.

Condition	Region noise	Triplet noise	What'sUp	COLA	Wino.	SC	SC++	MMVP	Avg.
CLIP-GBC (global)	-	-	41.90	27.90	17.10	77.60	57.60	28.89	41.83
GPS-CLIP	0%	0%	**47.38**	**30.64**	**18.35**	**83.29**	**65.62**	**33.33**	**46.44**
Region corruption only	20%	0%	46.00	29.50	17.30	82.90	62.93	28.33	44.49
Triplet corruption only	0%	20%	45.00	30.00	16.80	83.00	61.50	29.50	44.30
Joint corruption	20%	20%	45.20	29.10	16.60	81.32	61.35	28.15	43.62
Severe joint corruption	50%	50%	43.40	28.30	15.00	78.60	58.10	27.89	41.88

Moderate corruption. As shown in Table 23, corrupting either annotation stream individually at 20% retains a clear advantage over the global-only baseline, the region-only and triplet-only conditions obtain averages of 44.49 and 44.30, respectively, compared with 41.83 for CLIP-GBC. Joint 20% corruption produces a larger but still moderate decrease to 43.62. Together with the 10% joint-corruption result reported in Table 4, these results show that the structured advantage degrades gradually under moderate random annotation noise.

The per-benchmark behavior also reflects the different roles of the two supervision streams. At 20% corruption, region noise causes the larger MMVP decline ($33.33 \rightarrow 28.33$), consistent with MMVP emphasizing localized appearance, state, orientation, and structural cues. Triplet noise produces larger declines on What'sUp ($47.38 \rightarrow 45.00$), Winoground ($18.35 \rightarrow 16.80$), and SugarCrepe++ ($65.62 \rightarrow 61.50$), indicating greater sensitivity of these benchmarks to degraded relational supervision.

Severe-noise boundary. Under 50% joint corruption, the average falls to 41.88, effectively matching the global-only CLIP-GBC baseline of 41.83. Winoground and MMVP also fall below their corresponding global-only scores. Thus, although GPS-CLIP is tolerant to moderate random corruption, this stress test identifies a practical robustness boundary rather than suggesting annotation independence.

Our results suggest that the proposed formulation remains effective under moderate noise in such silver-standard supervision, and can benefit from improvements in annotation quality as these pipelines continue to advance.

E.7 Backbone Scaling to ViT-L/14

Table 24: **Effect of backbone capacity.** We report average accuracies (%) over compositional benchmarks (WhatsUp, COLA, Winoground, SugarCrepe, SugarCrepe++) and the fine-grained MMVP-VLM diagnostic, for CLIP and GPS-CLIP with ViT-B/32 and ViT-L/14 backbones. For each backbone block, the best method is in **bold**.

Backbone	Method	WhatsUp	COLA	Wino-ground	Sugar-Crepe	SugarCrepe ++	MMVP-VLM	Avg.
ViT-B/32	CLIP	41.8	22.8	16.4	72.9	54.6	27.42	39.32
ViT-B/32	***GPS-CLIP (Ours)***	**47.38**	**30.64**	**18.35**	**83.29**	**65.62**	**33.33**	**46.44**
ViT-L/14	CLIP	42.9	27.0	15.5	80.16	60.88	31.9	43.1
ViT-L/14	***GPS-CLIP (Ours)***	**48.9**	**32.86**	**19.5**	**86.17**	**66.18**	**36.3**	**48.32**

To assess whether our structured supervision continues to provide benefits as model capacity increases, we repeat our experiments with a **ViT-L/14** backbone using the same LoRA-based fine-tuning recipe and only light hyperparameter sanity checks (e.g., learning-rate scaling), without extensive tuning. Table 24 reports average accuracies across compositional benchmarks and MMVP-VLM for both CLIP and GPS-CLIP with **ViT-B/32** and **ViT-L/14** backbones.

For ViT-B/32, GPS-CLIP improves the OpenCLIP baseline from 39.32% to 46.44% average accuracy. For ViT-L/14, we again observe consistent gains, from 43.1% to 48.3%. These improvements are obtained under essentially the same training setup and without aggressive hyperparameter tuning for

the larger backbone, suggesting that they are a conservative estimate of the achievable performance. Overall, the results indicate that the benefits of region- and triplet-aware supervision persist and even amplify with increased backbone capacity, rather than being a small-model artifact.

E.8 SigLIP Compatibility and Design Discussion

The proposed structured objectives are directly compatible with sigmoid-based contrastive training. Region-span correspondences and IoU-based positive matching remain applicable, as the core contribution lies in defining grounded positive structure rather than relying on a specific normalization scheme.

Under a sigmoid formulation, the design can be interpreted as assigning positive labels to IoU-matched region pairs, negative labels to cross-image mismatches, and ignoring ambiguous same-image non-positive pairs. This avoids the need for denominator masking while preserving the intended supervision signal. The global alignment and relation-aware triplet terms remain unchanged, as they operate on global embeddings and do not depend on region-level normalization.

Conceptually, SigLIP and GPS-CLIP address complementary issues: sigmoid loss removes softmax-induced pairwise competition, while GPS-CLIP reduces incorrect supervision by enforcing grounded positive alignment. Empirically, we observe consistent gains when adding structured objectives to a SigLIP backbone, improving average accuracy from 44.34 to 48.60 (Table 5), mirroring improvements observed under InfoNCE training. These results indicate that the benefits of GPS-CLIP arise from structured alignment rather than the choice of contrastive loss.

SigLIP setup. For SigLIP experiments, we use a ViT-B/16 model pretrained on the WebLI dataset following the OpenCLIP implementation. Training follows the same fine-tuning protocol as our CLIP experiments, with the contrastive objective replaced by a sigmoid-based loss along with the design of region and relation losses. All other settings, including data, optimization, are kept identical unless otherwise specified.

SigLIP + GBC (global) and SigLIP + GBC + Structured Loss are independent matched runs from the same pretrained checkpoint, not sequential stages. Global-only fine-tuning improves five of six benchmarks (average 44.34 $\rightarrow$ 45.39), with SC++ the only decline (57.35 $\rightarrow$ 56.00), plausibly due to SC++'s strict requirement that both paraphrased positives outrank the negative. Adding the structured objectives reverses this, raising SC++ to 63.00 and the average to 48.60.

E.9 Region-Loss Component Ablation

Motivation. Region-level alignment is complicated by the presence of multiple overlapping regions that may describe shared semantic content. A conventional single-positive formulation can treat semantically compatible same-image regions as negatives, producing false-negative gradients. Our region objective addresses this ambiguity through three components: masking non-positive same-image pairs, promoting compatible overlaps to additional positives, and weighting those positives according to their IoU.

Design. We isolate these components using four controlled variants. The *unmasked single-positive* variant retains non-matching same-image pairs as negatives. The *masked single-positive* variant excludes all other same-image pairs from the denominator. The *masked uniform multi-positive* variant promotes compatible overlaps to equally weighted positives. Finally, the complete *masked IoU-weighted* objective weights these additional positives according to their spatial overlap. All variants use the same GBC supervision, backbone, global and relation objectives, optimizer, training schedule, and IoU threshold.

Component-level results. As shown in Table 25, same-image masking improves the six-benchmark average from 41.88 to 43.34 ($+1.46$), demonstrating the cost of treating ambiguous same-image pairs as negatives. Promoting compatible overlaps to uniformly weighted positives provides a further improvement from 43.34 to 43.70 ($+0.36$). Calibrating their contributions by IoU produces the largest additional gain, increasing the average from 43.70 to 46.44 ($+2.74$). Together, the three components improve the average by 4.56 points over the unmasked single-positive formulation.

Per-benchmark results. Table 26 reports the per-benchmark comparison between the masked single-positive and complete IoU-weighted formulations. The complete IoU-weighted multi-positive

Table 25: **Controlled decomposition of the region objective.** Each row introduces one additional component while holding the remaining training configuration fixed.

Region-loss variant	Same-image masking	Overlap positives	Weighting	Avg.
Unmasked single-positive	No	No	-	41.88
Masked single-positive	Yes	No	-	43.34
Masked uniform multi-positive	Yes	Yes	Uniform	43.70
Masked IoU-weighted	Yes	Yes	IoU	**46.44**

Table 26: **Effect of multi-positive region loss.** We compare single-positive and multi-positive region formulations under the same GPS-CLIP configuration.

Model	What'sUp	COLA	Wino	SC	SC++	MMVP	Avg.
GPS-CLIP w/ single-pos. region loss	43.00	30.00	16.00	80.78	62.25	28.00	43.34
GPS-CLIP w/ multi-pos. region loss	**47.38**	**30.64**	**18.35**	**83.29**	**65.62**	**33.33**	**46.44**

formulation improves all six benchmarks relative to the masked single-positive variant, increasing the average from 43.34 to 46.44 ($+3.10$). The largest improvements occur on What'sUp, SugarCrepe++, and MMVP, while the remaining benchmarks also improve.

Discussion. The three components play distinct roles. Masking prevents ambiguous same-image pairs from generating false-negative gradients, multi-positive assignment determines which compatible overlaps contribute positively, and IoU weighting calibrates their contributions according to spatial compatibility. The relatively small gain from uniform promotion ($+0.36$), compared with the larger gain from IoU weighting ($+2.74$), shows that treating all overlaps identically is insufficient. The improvement therefore arises from the complete overlap-aware assignment rule rather than from region-text alignment or multi-positive promotion alone.

E.10 Directional Relation Diagnostics

Scope of the relation objective. The triplet objective aligns a full-image embedding with textual subject-relation-object triplets. In Eq. 4 (Sec. 3.6), same-image masking is applied only in the image→triplet direction; no masking is needed in the triplet→image direction, since each image appears exactly once among the N candidates and there are no same-image non-positives. It does not construct explicitly localized subject and object representations or a predicate-specific visual embedding. Instead, it provides an ordered, scene-level relational inductive bias to the same global image representation used by the standard dual-encoder inference interface. The region and triplet objectives therefore play complementary roles; the region objective aligns entities and attributes with localized visual evidence, while the triplet objective introduces ordered relational supervision over (s, r, o) configurations.

Why order sensitivity can emerge. Each triplet is encoded in its observed order. Consequently, (s, r, o) and its role-swapped counterpart (o, r, s) generally produce different text embeddings despite containing the same entities and predicate. The contrastive objective pulls the image embedding toward the ordering observed in the scene and away from triplets associated with other images. This can induce sensitivity to relational order. However, the objective does not explicitly contrast every correct triplet against its role-swapped counterpart, its directional supervision is therefore indirect and dependent on the relational patterns present in the training data.

Direct role-swap probes. We evaluate whether the learned image-text similarity prefers the correct triplet ordering over a counterfactual obtained by exchanging the subject and object while keeping the predicate fixed. For an image I, correct triplet (s, r, o), and swapped triplet (o, r, s), we define the role-swap margin as

$$\Delta_{\text{swap}} = \text{sim}\big(v_I, t_{(s,r,o)}\big) - \text{sim}\big(v_I, t_{(o,r,s)}\big) .$$

Role-swap accuracy is the fraction of examples for which $\Delta_{\text{swap}} > 0$.

Table 27: **Direct role-swap diagnostics on ARO Visual Genome Relations.** Accuracy measures how often the image prefers the correct ordered triplet over its subject-object-swapped counterpart. A positive mean margin indicates an average preference for the correct order.

	Asymmetric relations		Spatial only	
Model	**Role-swap accuracy**	**Mean margin**	**Role-swap accuracy**	**Mean margin**
CLIP-GBC (global)	46.74%	−0.001066	44.17%	−0.001489
CLIP-GBC + Region	53.58%	0.000333	56.13%	0.000525
CLIP-GBC + Triplet	57.11%	0.001386	55.97%	**0.001937**
GPS-CLIP	**57.40%**	**0.002060**	**57.53%**	0.001587

We construct the probes from the ARO Visual Genome Relations split. Subset selection depends only on predicate asymmetry and source frequency and does not use predictions from any evaluated model. Because the two candidate triplets contain exactly the same entities and predicate, differing only in role order, the probe removes a simple bag-of-words explanation.

Asymmetric-relation probe. The first probe contains 5,631 pairs covering 11 asymmetric predicates: *on*, *wearing*, *in*, *in front of*, *behind*, *above*, *below*, *on top of*, *sitting on*, *holding*, and *under*. Approximately symmetric and coordinate predicates are excluded.

Spatial-only probe. The second probe contains 4,365 pairs covering eight spatial predicates: *on*, *in*, *in front of*, *behind*, *above*, *below*, *on top of*, and *under*. Agentive predicates are removed to reduce the possibility of solving the task using subject-object plausibility alone.

Results. As shown in Table 27, global-only GBC fine-tuning performs below chance on both probes and produces negative mean margins, indicating an average preference for the swapped ordering. On the broader asymmetric-relation probe, adding the triplet objective increases accuracy from 46.74% to 57.11% and changes the mean margin from -0.001066 to 0.001386. This is a larger single-objective improvement than that obtained with region supervision and supports the intended role of the triplet objective in learning ordered subject-relation-object compatibility.

On the more restrictive spatial-only probe, the region-only and triplet-only variants achieve similar role-swap accuracy, while the complete model obtains the highest accuracy of 57.53%. The triplet-only variant obtains the largest mean margin on this subset, so the complete model does not dominate every reported statistic. Nevertheless, the consistent transition from negative margins under global-only training to positive margins under structured supervision indicates improved sensitivity to role order and spatial direction.

Interpretation. These diagnostics support the claim that GPS-CLIP joint image-text scoring becomes sensitive to relational order. They do not establish explicit predicate localization; the visual input to the triplet objective remains the full-image embedding. We therefore characterize the relation term as an *ordered, scene-level relational inductive bias*, rather than as a predicate-specific grounding mechanism.

E.11 Controlled Baselines on the GBC Subset

Motivation. To distinguish the contribution of the GBC training data, access to structured annotations, and our objective design, we construct two controlled baselines on the same filtered GBC subset used by GPS-CLIP. The first tests whether simply applying standard contrastive learning to scene-graph annotations while removing the proposed structured inductive biases is sufficient, while the second tests whether caption-side hard-negative training can account for the observed gains.

Naive structured baseline. This baseline uses the same GBC bounding boxes and relation triplets as GPS-CLIP, but trains region-span and image-triplet pairs using standard single-positive InfoNCE objectives. It does not apply IoU-based multi-positive grouping, IoU weighting, or intra-image masking. Consequently, same-image non-matching regions enter the denominator as negatives. This baseline therefore preserves access to structured annotations while removing the proposed alignment mechanisms.

NegCLIP-GBC baseline. We additionally train NegCLIP-GBC by applying the original NegCLIP caption-side hard-negative generation pipeline and training loss to the global captions from the same GBC subset. NegCLIP-GBC and GPS-CLIP use the same training images and global captions, OpenCLIP ViT-B/32 initialization pretrained on LAION-2B, LoRA configuration, effective batch size, optimizer, learning-rate schedule, number of epochs, and random seed. Each method retains its native supervision: NegCLIP-GBC uses generated caption-side hard negatives, whereas GPS-CLIP uses grounded region- and relation-level positives.

Results. As shown in Table 28, the naive structured baseline provides only a marginal improvement over global GBC fine-tuning ($41.83 \rightarrow 42.10$). NegCLIP-GBC reaches an average of 41.31, whereas GPS-CLIP achieves 46.44, outperforming the matched NegCLIP-GBC control by 5.13 points and obtaining the best result on every reported benchmark.

Analysis. The two controls reveal different limitations. The naive structured baseline degrades performance relative to CLIP-GBC on What'sUp, Winoground, and SugarCrepe, consistent with false-negative pressure introduced when semantically related same-image regions are treated as negatives. GPS-CLIP instead masks same-image non-positives and incorporates overlapping regions as additional positives, reinforcing their shared semantic content rather than pushing it apart.

NegCLIP-GBC improves the caption-perturbation-sensitive benchmarks What'sUp, SugarCrepe, and SugarCrepe++ relative to CLIP-GBC, indicating that generated hard negatives provide a useful training signal. However, these gains are accompanied by lower performance on COLA, Winoground, and MMVP. Grounded positive supervision therefore produces broader improvements than caption-side hard-negative training under the matched GBC setup.

The specialization of NegCLIP-GBC is also reflected in zero-shot recognition. Its ImageNetV2 Top-1 accuracy decreases from 55.96 for pretrained CLIP to 52.03 (Top-5: $83.41 \rightarrow 80.06$), whereas GPS-CLIP reaches 58.11 Top-1 and 83.67 Top-5. Thus, although caption-side hard-negative training improves sensitivity to selected linguistic perturbations, these gains do not translate into broader zero-shot recognition. This pattern is consistent with the modality-asymmetry discussion in App. B.2, caption-side negatives emphasize linguistic distinctions, whereas grounded positive supervision additionally reinforces their correspondence with localized visual content.

Discussion. Together, these controls show that the gains of GPS-CLIP are not explained solely by fine-tuning on GBC or by access to structured annotations. They instead depend on how the structure is incorporated, GPS-CLIP uses bounding boxes and relation triplets to define grounded positive alignments rather than relying on single-positive region matching or text-side hard-negative construction. This distinguishes our approach from prior scene-graph-based methods that primarily leverage structure for text-side supervision or hard-negative construction, rather than image-side positive alignment in a shared embedding space. Notably, GPS-CLIP outperforms NegCLIP-GBC on all six compositional benchmarks, including SugarCrepe and SugarCrepe++, where hard-negative training is expected to be particularly effective. The controlled conclusion is therefore not that hard negatives provide no useful signal, but that, under matched GBC data, initialization, and optimization settings, the NegCLIP objective does not reproduce the broader gains obtained through grounded positive supervision. Together with the global-only and naive structured controls, this indicates that GPS-CLIP's advantage is explained neither by the GBC data nor by access to structured annotations alone, but by using that structure to construct grounded positive constraints.

Table 28: **Controlled comparisons on the same filtered GBC subset.** All fine-tuned models use matched training images, global captions, initialization, and optimization settings. The naive structured baseline and GPS-CLIP additionally use the same GBC boxes and relation triplets, but differ in how this supervision is incorporated.

Model	What'sUp	COLA	Wino	SC	SC++	MMVP	Avg.
CLIP-GBC (global)	41.90	27.90	17.10	77.60	57.60	28.89	41.83
CLIP + Naive structured baseline	40.00	30.00	16.58	76.00	60.00	30.00	42.10
NegCLIP-GBC	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**

F Discussion on Scalability to Large-Scale Datasets

A potential concern regarding GPS-CLIP is its reliance on structured scene-graph supervision, which is traditionally more expensive to obtain than raw image-text pairs. We emphasize, however, that GPS-CLIP is designed as a *post-pretraining compositional alignment stage*, and is compatible with scalable, automatically generated supervision. We outline four factors that support its scalability:

(i) **Data Efficiency.** Our results indicate that structured supervision provides a strong inductive bias for compositional reasoning when used appropriately. While simply exposing the model to structured annotations yields only a modest improvement (41.83 $\rightarrow$ 42.10), incorporating our structured alignment objectives leads to a substantially larger gain (41.83 $\rightarrow$ 46.44). This suggests that improvements arise not merely from access to richer data, but from how that structure is leveraged during training. Compared to hard-negative approaches, which often rely on large numbers of synthetic examples, our formulation achieves stronger gains through targeted positive alignment.

(ii) **Scalable Automated Extraction.** Modern pipelines such as GBC demonstrate that scene-graph annotations can be generated automatically at scale using open-vocabulary detectors and MLLM-based caption parsing, without manual labeling. In this setting, scene graphs act as *silver-standard supervision*, trading perfect accuracy for scalability. Our robustness analysis (App. E.6) shows that GPS-CLIP remains effective under moderate noise, indicating compatibility with such automatically generated annotations.

(iii) **Lightweight Integration into Training Pipelines.** GPS-CLIP does not require architectural modifications or changes to inference. Structured supervision is used only during training, and the method can be incorporated as an auxiliary objective alongside standard contrastive learning. In practice, this enables mixed-objective training (e.g., combining unstructured image-text pairs with a subset of structured samples), preserving broad visual knowledge while improving compositional alignment.

(iv) **No Additional Data Generation Overhead.** Unlike hard-negative mining methods that require explicit generation of counterfactual captions or images, GPS-CLIP operates directly on existing annotations. This avoids introducing additional synthetic data pipelines beyond those already used for large-scale dataset construction, making the approach simpler and more computationally efficient in practice.

Extension to mixed-corpus training. A natural large-scale extension is mixed-corpus training, the global contrastive objective could operate on all image-text pairs, while the region and triplet objectives are applied only to samples with sufficiently reliable structured annotations. Our current setup already separates the two by stage, broad unstructured pretraining (the OpenCLIP LAION-2B model) followed by a small structured alignment stage on a grounded subset rather than a direct test of web-scale training over mixed structured and unstructured samples.

Taken together, these properties position GPS-CLIP not as a replacement for web-scale pretraining, but as a practical and scalable fine-tuning stage that leverages increasingly available structured supervision to improve compositional reasoning.

G Qualitative Examples and Failure Modes

G.1 Successful dense-caption alignment.

Figures 4, 6 and 7 show examples where GPS-CLIP assigns the highest similarity to dense captions containing multiple entities, attributes, and relations.

G.2 Qualitative Near-Miss Cases

Although GPS-CLIP improves compositional alignment across benchmarks, it does not uniformly assign the highest similarity for every dense caption. Fig. 8 shows two representative near-miss cases where GPS-CLIP remains competitive with strong baselines but is slightly out-scored by another method.

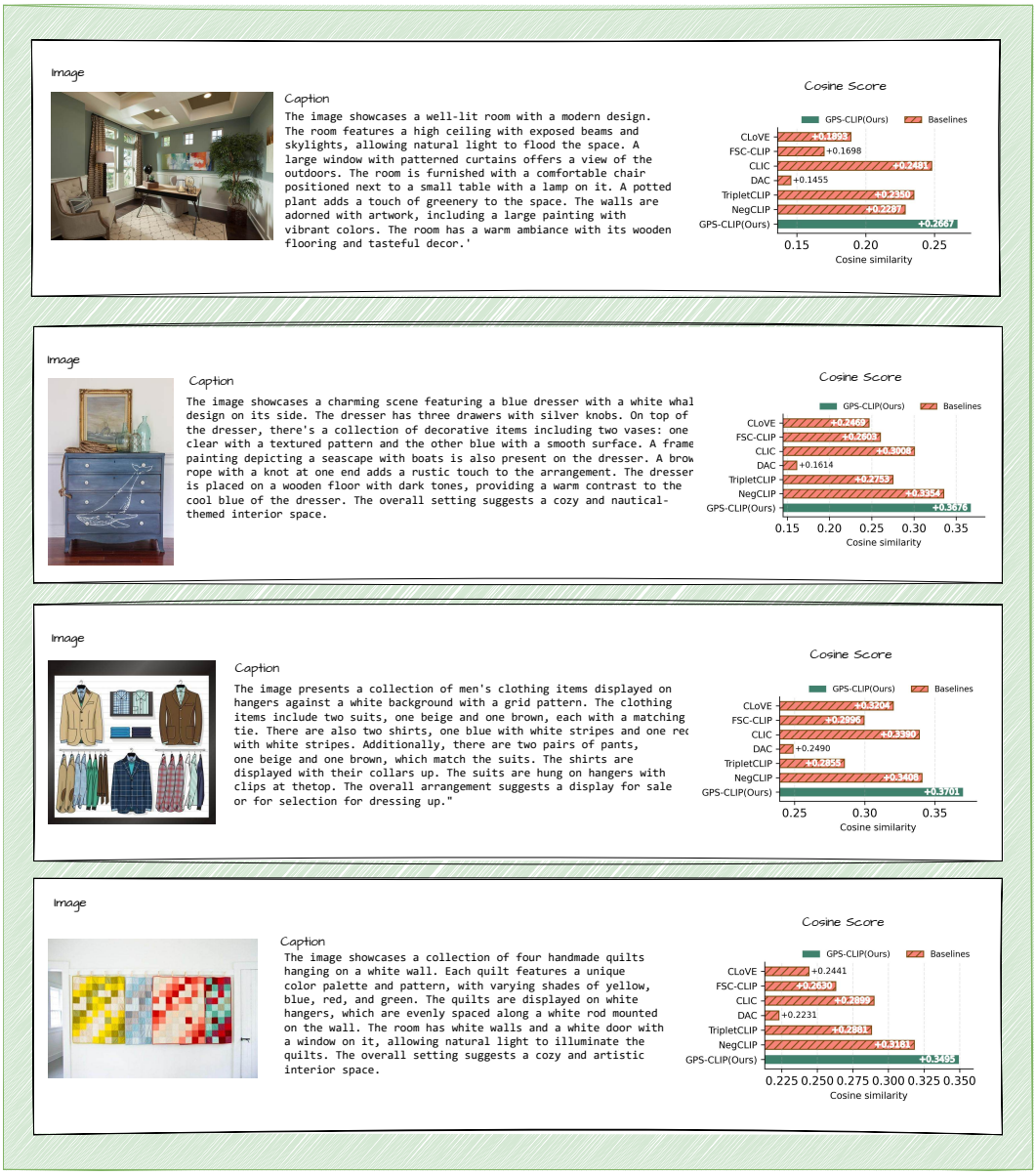


Figure 6: **Dense caption alignment across contrastive objectives.** Each row shows an image, its dense caption, and the image–caption cosine similarity for GPS-CLIP and baselines spanning hard-negative-based and compositional alignment methods. The examples cover indoor scenes with multiple entities, attributes, and relations. GPS-CLIP yields the highest similarity in each case, suggesting that region- and relation-aware positive supervision improves alignment with complex compositional descriptions.

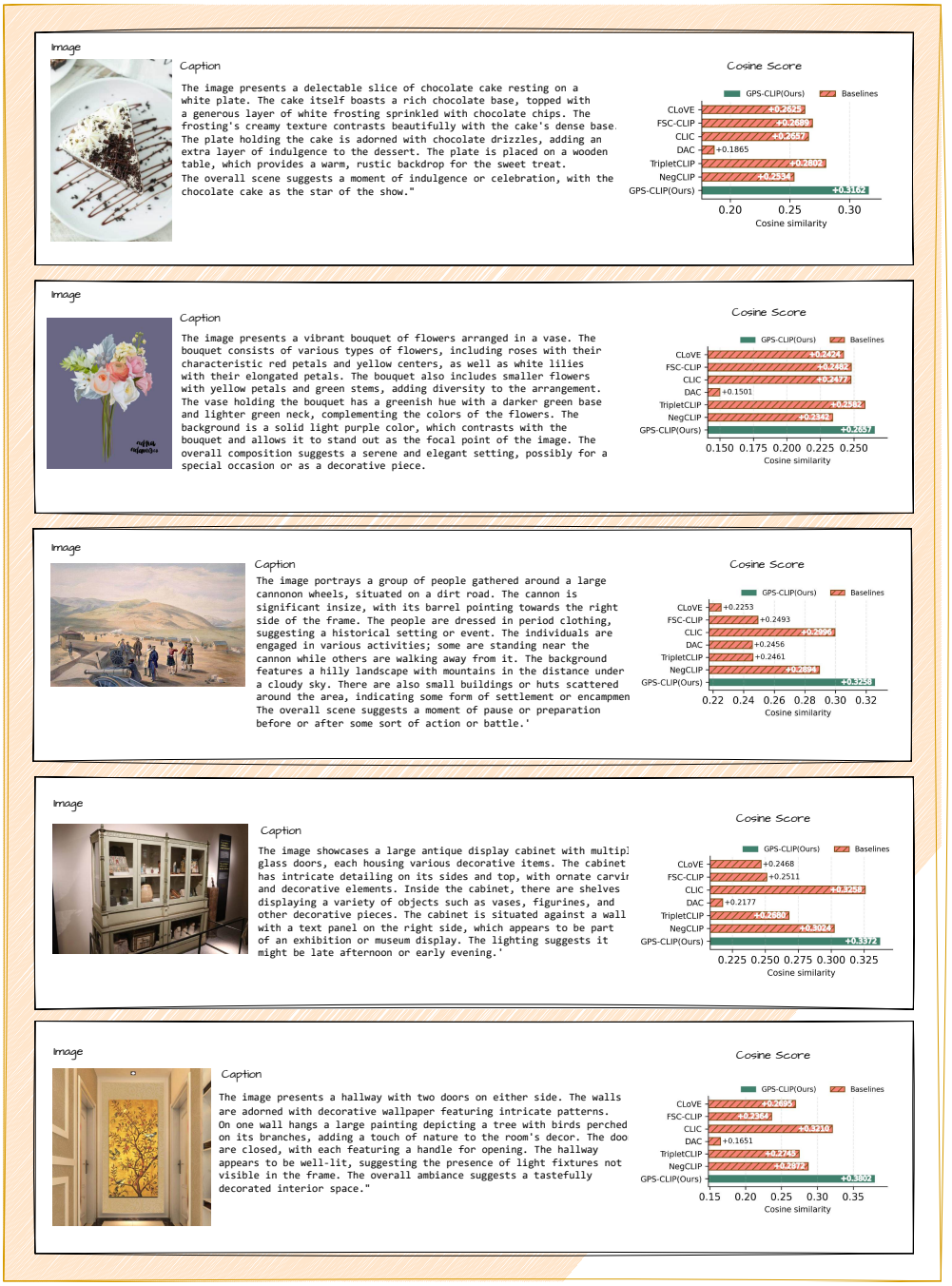


Figure 7: **Additional dense caption alignment examples** (continued from Fig. 6), including outdoor and cluttered scenes. GPS-CLIP again assigns the highest similarity. The last row is also shown in Fig. 4.



Figure 8: **Representative near-miss cases.** GPS-CLIP generally improves dense-caption alignment, but does not uniformly assign the highest similarity across all examples. Top: a cluttered multi-actor tavern scene with repeated entities, mugs, barrels, and overlapping social interactions, where NegCLIP assigns a higher similarity. Bottom: an attribute-rich decorative bouquet scene with fine-grained visual details, where CLIC slightly out-scores GPS-CLIP. These cases suggest that positive grounded alignment improves object-attribute and relation binding overall, but crowded multi-entity scenes and highly detailed dense descriptions can still favor hard-negative or caption-augmentation baselines.

In the tavern example, the scene contains many repeated entities and overlapping social interactions, making it difficult to assign a single clean set of region–span and relation-level correspondences. GPS-CLIP remains stronger than most baselines but scores below NegCLIP, suggesting that dense multi-actor settings may still benefit from explicit hard-negative pressure. In the bouquet example, GPS-CLIP is nearly tied with the best method but is slightly below CLIC, indicating that caption-augmentation methods can sometimes better match highly detailed decorative descriptions. Overall, these examples clarify the scope of GPS-CLIP: positive grounded supervision improves compositional binding, but does not replace specialized supervision for exhaustive dense captioning, rare spatial distinctions, or highly fine-grained localization.

H Broader Impact

GPS-CLIP is a methodological contribution for improving compositional alignment in CLIP-style vision-language models. Its main expected impact is to make vision-language representations more reliable for fine-grained visual understanding, especially object-attribute binding, relation-aware retrieval, and dense image-text matching. By encouraging models to align language with grounded visual evidence rather than relying primarily on shortcut correlations or synthetic hard negatives, GPS-CLIP may help reduce compositional mismatches in downstream retrieval and recognition systems.

The method preserves the standard dual-encoder inference interface: scene graphs, regions, and triplets are used only during fine-tuning, with no additional inference-time modules or structured inputs. This makes GPS-CLIP a lightweight post-pretraining alignment stage rather than a deployed system or a new data-generating model. As with any vision-language representation, downstream applications should still account for biases inherited from the pretrained backbone and training data, especially in sensitive recognition or retrieval settings.