# DEPTH-WISE SPECIALIZATION IN VISION-LANGUAGE MODELS FOR COMPOSITIONAL ADAPTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Contrastive vision–language models such as CLIP and SigLIP achieve strong global image–text alignment, yet often struggle with fine-grained compositional distinctions such as modifier sensitivity and visually grounded relations. Prior work improves compositionality through stronger supervision, caption enrichment, or pooling modifications, but largely overlooks how compositional cues evolve across encoder depth. We reveal complementary depth specialization: noun-phrase selectivity strengthens toward later layers, whereas modifier- and relation-sensitive distinctions progressively attenuate. Depth-wise probes of broad noun-phrase selectivity, same-head modifier sensitivity, and relation-family discrimination consistently support this pattern. Motivated by this finding, we introduce RELIP, a depth-asymmetric fine-tuning method for contrastive dual encoders. RELIP retains each backbone's native global image–caption objective while using shallow patch representations as compositional evidence and a detached late relevance map to emphasize attribute–entity and grounded-relation cues that attenuate with depth. The auxiliary branch is training-only and leaves the original inference architecture unchanged. Under the same CC300K budget, RELIP improves both CLIP and SigLIP; with SigLIP, it reaches $86.1$ on SugarCrepe and $66.9$ on SugarCrepe++, outperforming $C^2$LIP by $1.4$ and $2.2$ points while preserving strong retrieval and zero-shot performance. These results show that effective compositional adaptation depends not only on *what* supervision is provided, but also on *where* and *how* it is applied.

## 1 INTRODUCTION

Contrastive vision–language models such as CLIP Radford et al. (2021) and SigLIP Zhai et al. (2023) provide strong global image–text alignment and zero-shot transfer, but remain limited in compositional understanding. They may recognize the objects in a caption while failing to bind modifiers to the correct entities Yuksekgonul et al. (2023); Hsieh et al. (2023); Dumpala et al. (2024), distinguish visually grounded relations Kargi et al. (2026); Yuksekgonul et al. (2023); Thrush et al. (2022), or preserve phrase specificity within longer descriptions Yuksekgonul et al. (2023); Kamath et al. (2023). Improving these abilities without degrading the broad transferability of pretrained models remains an important challenge Peleg et al. (2025b); Pham et al. (2026).

Prior work primarily changes the supervision or the final readout, using synthetic hard negatives Yuksekgonul et al. (2023); Singh et al. (2024); Zhang et al. (2024); Doveh et al. (2023a;b); Patel et al. (2024), richer captions Zheng et al. (2024), or shorter concept-level phrases Pham et al. (2026). These studies establish the value of fine-grained supervision, but leave a complementary question underexplored: *is the same compositional cue equally accessible at every visual depth?*

We study this question using phrase-conditioned patch relevance probes that separate broad noun-phrase selectivity from fine-grained compositional discrimination. Late representations increasingly concentrate relevance on dominant noun phrases, whereas same-head modifier-binding and grounded-relation contrast margins attenuate from shallow to middle to late layers. Figure 1 qualitatively previews this attenuation. This pattern goes beyond the generic distinction between low- and high-level features: noun-phrase selection becomes increasingly concentrated, while compositional specificity progressively weakens across depth.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




Figure 1: Depth-wise specialization of compositional evidence. For representative attribute and relation pairs, shallow residuals preserve clearer phrase-specific evidence than late residuals. Warm and cool regions favor phrases $A$ and $B$, respectively; this asymmetry motivates RELIP.

Motivated by this finding, we introduce RELIP, a depth-asymmetric fine-tuning objective for contrastive dual encoders. RELIP uses the normalized shallow–late relevance difference to spatially weight shallow patch evidence, while treating the detached late map as a semantic reference. The native global image–caption objective preserves global alignment, and the auxiliary branch is removed at inference.

Under controlled CC300K fine-tuning, RELIP consistently improves both CLIP and SigLIP, including over the composition-aware $C^2$LIP baseline, while preserving retrieval and zero-shot classification performance.

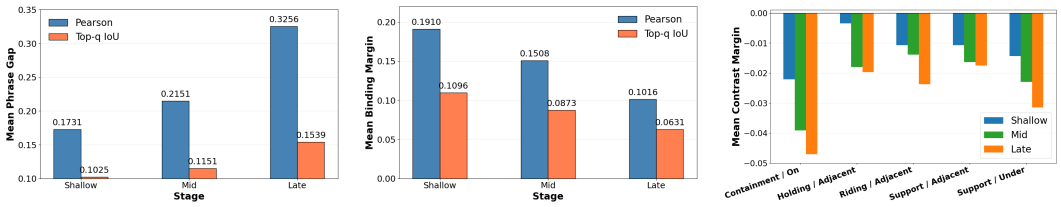
Our contributions are:

1. We reveal complementary depth specialization: noun-phrase selectivity strengthens with depth, while modifier- and relation-sensitive distinctions attenuate.
2. We propose RELIP, which uses a detached late reference to spatially weight depth-attenuated shallow compositional evidence without modifying inference.
3. Experiments on CLIP and SigLIP demonstrate consistent compositional gains, while controlled ablations isolate the effects of phrase source, supervision depth, and differential weighting.

## 2 Related Work

**Compositionality Failures in Vision-Language Models.** Despite strong global retrieval and zero-shot capabilities, contrastive vision-language models fundamentally struggle with compositional understanding, particularly in attribute binding, relation discrimination, word-order sensitivity, and numeracy Burapacheep et al. (2024); Lewis et al. (2024); Gurung et al. (2025); Paiss et al. (2023). Early benchmarks such as Winoground Thrush et al. (2022), VALSE Parcalabescu et al. (2022), ARO Yuksekgonul et al. (2023), and What'sUp Kamath et al. (2023) broadly exposed these pervasive weaknesses. More importantly, rigorous evaluation protocols like SugarCrepe confirm that these failures are inherent representational bottlenecks rather than exploitable dataset artifacts, demonstrating that models often rely on bag-of-words heuristics rather than true structured grounding Hsieh et al. (2023).

**Supervision and Data-Centric Remedies.** A major line of work improves compositionality by changing the training signal. Hard-negative methods perturb captions, images, or both to make contrastive learning more sensitive to attributes and relations Yuksekgonul et al. (2023); Patel et al. (2024); Singh et al. (2024); Zhang et al. (2024). Component-supervised methods decompose captions into entity and relation units to penalize partial mismatches Kargi et al. (2026), while caption-centric




(a) Mean phrase gap (b) Mean modifier-binding margin (c) Relation-family contrast margin

Figure 2: Depth-wise specialization of phrase-conditioned visual relevance. (a) Broad noun-phrase selectivity, measured by the mean phrase gap, increases from shallow to middle to late representations. (b) Same-head modifier-binding margins progressively decrease across depth under both Pearson correlation and top-$q$ IoU. (c) Relation contrast margins exhibit the same ordered attenuation across all five relation families; larger, i.e., less negative, values indicate better discrimination.

methods use rewriting, dense captions, or better-aligned texts to provide richer positives Doveh et al. (2023b;a); Zheng et al. (2024); Liu et al. (2024). These works show that better supervision matters, but stronger supervision alone is not sufficient: hard-negative gains can be overstated on artifact-prone benchmarks Hsieh et al. (2023). RELIP builds on phrase-level supervision but asks where in the visual encoder those cues are best preserved.

**Architectural Bottlenecks, Fine-tuning, and Our Position.** Recent work has examined architectural and representational bottlenecks in contrastive vision-language models. Analyses of modality gaps, object bias, and information imbalance suggest that global contrastive representations can favor salient object-level summaries over finer-grained factors Liang et al. (2022); Schrodi et al. (2025). C$^2$LIP identifies long captions and final global pooling as key obstacles Pham et al. (2026); Sparo and FLAIR modify readout or pooling mechanisms Vani et al. (2024); Xiao et al. (2024), while READ targets the text encoder Kwon (2025). Recent fine-tuning work also shows that compositional gains do not always transfer to stronger semantic evaluation or retrieval Peleg et al. (2025a). These works largely focus on the encoder endpoint, readout, or text side. RELIP instead studies how compositional cues evolve within the visual encoder and uses a shallow-late differential objective for attribute-entity and grounded-relation phrases.

## 3 Empirical Motivation

We first ask whether phrase-conditioned visual relevance changes systematically across encoder depth. To answer this, we probe three complementary behaviors: broad noun-phrase selectivity, modifier-sensitive binding, and relation-sensitive discrimination. As previewed in Figure 1, shallow representations retain more phrase-specific compositional evidence, whereas late representations favor stronger but coarser noun-phrase selectivity. The goal of this section is to verify that this pattern holds consistently across datasets and probe types, thereby motivating RELIP's asymmetric use of shallow evidence and a detached late reference.

### 3.1 Experimental Setup for Depth-wise Probing

Our analysis uses all visual blocks of a pretrained CLIP ViT-L/14 and the MSCOCO validation set Lin et al. (2014), with one fixed caption per image. At visual block $l$, we extract the pre-attention normalized patch tokens from its $\mathrm{LN}_1$ output, discard the CLS token, and apply the shared visual projection followed by L2 normalization. For a text query $q$ (full caption or caption phrase), we similarly normalize its final text embedding and compute

$$r^{(l)}_{I,q,m} = \cos\left(z^{(l)}_{I,m}, e_q\right), \tag{1}$$

where $m$ indexes spatial patches. We denote the resulting relevance map by $R^{(l)}(I, q)$.

We use Qwen3-4B-Instruct-2507 Team (2025) only to identify entity, attribute–entity, and grounded-relation spans already present in the original captions. For caption $c$ and phrase $p$, $s_l(c, p)$ measures

the alignment between their relevance maps using Pearson correlation or top-$q$ IoU, which respectively capture full-map agreement and overlap among the most activated patches; we use $q = 0.3$.

We use three complementary probes. The noun-phrase probe measures the mean absolute pairwise difference among the alignment scores of entity and attribute–entity phrases within a caption, with a larger phrase gap indicating more uneven phrase selectivity. The modifier-sensitive probe compares an attribute–entity phrase with its bare head noun, such as *red frisbee* versus *frisbee*, to measure modifier-induced spatial specificity. The relation probe measures the margin between a grounded target relation and its hardest confusable alternative. Each probe is computed independently at every block, after which the layer-wise metric values are averaged within shallow stages (layers 0–7), middle (8–15) and late (16–23); relevance maps are not averaged across layers.

### 3.2 Noun-Phrase Selectivity Strengthens with Depth

We first examine how selectively different caption-relevant noun phrases are represented across depth. For each caption, we pair its full-caption relevance map with the entity and attribute–entity phrases extracted from that caption. The phrase gap measures the average pairwise difference among the resulting caption–phrase alignment scores; a larger value therefore indicates that the caption-conditioned representation aligns more strongly with some noun phrases than with others.

As shown in Figure 2a, the mean phrase gap increases consistently from shallow to middle to late representations. Under Pearson correlation, it rises from $0.1731$ to $0.2151$ and then to $0.3256$. The corresponding top-$q$ IoU values increase from $0.1025$ to $0.1151$ and $0.1539$. Thus, the stronger late-layer selectivity is not produced by an abrupt difference between two manually chosen endpoints; the middle stage lies between the shallow and late stages under both metrics. The same endpoint trend is observed at the sample level. As reported in Appendix, $71.7\%$ of samples have a larger late-than-shallow phrase gap under Pearson correlation, and $73.5\%$ do so under top-$q$ IoU. These results show that caption-conditioned alignment becomes increasingly uneven across noun phrases with depth. However, the phrase gap does not determine whether closely related phrases remain distinguishable. We next test this question using controlled same-head and grounded-relation comparisons.

### 3.3 Fine-grained Compositional Discrimination

We examine two controlled forms of compositional discrimination: modifier-sensitive specificity and grounded-relation contrasts. Figure 1 provides a qualitative preview of both patterns: shallow residuals preserve clearer phrase-specific evidence for representative attribute and relation pairs, whereas late residuals are weaker and more diffuse. We next test whether this attenuation holds systematically from shallow to middle to late depth.

#### 3.3.1 Modifier-sensitive Binding

We isolate the visual information contributed by a modifier using same-head phrase pairs, such as *red frisbee* and *frisbee*. For a modifier-bearing phrase $p$ and its bare head noun $h$, we define the binding margin at layer $l$ as

$$B_l(p, h) = s_l(c, p) - s_l(c, h), \tag{2}$$

where $s_l(c, \cdot)$ denotes the alignment between the full-caption relevance map and the corresponding phrase relevance map at layer $l$. A larger margin indicates that the representation better preserves the spatial specificity introduced by the modifier.

As shown in Figure 2b, the mean binding margin decreases from shallow to middle to late representations: $0.1910$, $0.1508$, and $0.1016$ under Pearson correlation, and $0.1096$, $0.0873$, and $0.0631$ under top-$q$ IoU. The shallow–late differences are also consistent at the pair level: $68.7\%$ of phrase pairs yield larger shallow margins under Pearson correlation, and $61.8\%$ do so under top-$q$ IoU. Both endpoint differences are significant under image-level paired Wilcoxon tests ($p = 5.88 \times 10^{-5}$ and $p = 2.67 \times 10^{-3}$), while the adjacent comparisons jointly support ordered attenuation under Pearson ($p_{\mathrm{IUT}} = 3.20 \times 10^{-15}$) and top-$q$ IoU ($p_{\mathrm{IUT}} = 8.10 \times 10^{-9}$); endpoint distributions are provided in the Appendix.

These findings complement the noun-phrase selectivity analysis. Although later representations become increasingly selective among broad noun phrases, the additional specificity introduced by a

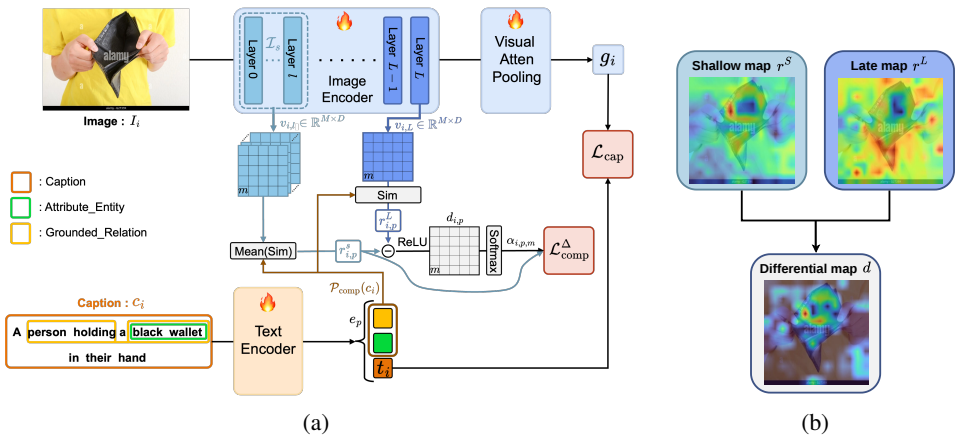



Figure 3: Overview of RELIP. (a) SigLIP instantiation, where a detached late relevance map guides differential weighting of shallow compositional evidence. For CLIP, visual attention pooling is replaced by the backbone's native CLS-token global readout; the auxiliary branch is otherwise unchanged. (b) Shallow, late, and differential relevance maps used by the training-only objective.

modifier is progressively attenuated with depth. Thus, stronger late-layer selectivity does not imply uniformly better phrase understanding; it increasingly emphasizes the shared head-entity semantics over the modifier-sensitive distinction.

### 3.3.2 Relation-Sensitive Discrimination

We next examine whether the depth-dependent attenuation observed for modifier binding also extends to visually grounded relations. We construct a curated relation-family benchmark in which each episode contrasts a target relation with nearby confusable alternatives that largely share the same entities. The five relation-family contrasts are containment versus on, holding versus adjacency, riding versus adjacency, support versus adjacency, and support versus under.

For each episode $e$, we define the depth-specific contrast margin as $C_l(e) = s_l^+(e) - \max_{n \in \mathcal{N}_e} s_{l,n}^-(e)$, where $\mathcal{N}_e$ contains the confusable alternatives and a larger margin indicates stronger separation from the hardest alternative. As shown in Figure 2c, relation-sensitive discrimination weakens with depth. Averaged across the five relation families, the contrast margin decreases from $-0.0101$ in shallow representations to $-0.0210$ in middle and $-0.0300$ in late representations, with all family-level means following $C_{\text{shallow}} > C_{\text{middle}} > C_{\text{late}}$. The endpoint difference is particularly clear: shallow representations yield larger margins than late representations in $62.0\%$ of episodes, with mean and median paired increases of $0.0133$ and $0.0144$, respectively ($p = 2.42 \times 10^{-8}$; one-sided paired Wilcoxon test). The intermediate stage lies between the endpoints, and the adjacent comparisons jointly support the ordered attenuation ($p_{\text{IUT}} = 5 \times 10^{-4}$). Although the episode-level shifts are modest, their consistent aggregate direction indicates that target relations become less separable from their closest alternatives at greater depth.

Together, the probes reveal complementary depth specialization: late representations show stronger noun-phrase selectivity, whereas shallow representations retain larger modifier- and relation-sensitive margins. This asymmetry motivates RELIP, which uses the late representation to identify depth-attenuated cues and supervises the corresponding shallow patch evidence.

## 4 Proposed Method

We propose **RELIP**, a depth-asymmetric compositional fine-tuning method motivated by the specialization identified in empirical motivation section. Because shallow representations better preserve modifier- and relation-sensitive distinctions, while late representations exhibit stronger noun-phrase selectivity, RELIP uses shallow patch tokens as trainable evidence and a detached late relevance map as a semantic reference. As shown in Figure 3, RELIP retains each backbone's native global

image–caption objective and adds the same training-only differential objective to CLIP ViT-B/16 and SigLIP ViT-B/16 Zhai et al. (2023).

### 4.1 Shallow–Late Differential Evidence

Given an image–caption pair $(I_i, c_i)$, we extract a phrase set $\mathcal{P}_{\text{comp}}(c_i)$ containing filtered attribute–entity and grounded-relation spans. Grounded relations include their visual arguments rather than isolated relation words.

For phrase $p$, let $e_p$ be its L2-normalized text embedding. At each selected visual block, we apply the backbone's visual projection and L2-normalize the patch tokens $v_{i,l,m}$. The shallow and late phrase-relevance maps are

$$\begin{aligned} r^s_{i,p,m} &= \frac{1}{|\mathcal{I}_s|} \sum_{l \in \mathcal{I}_s} \cos(v_{i,l,m}, e_p), \\ r^L_{i,p,m} &= \cos(v_{i,L,m}, e_p). \end{aligned} \tag{3}$$

To remove layer-dependent scale and offset differences, we z-normalize each map over spatial patches. The differential evidence and spatial weights are

$$\begin{aligned} d_{i,p,m} &= \mathrm{ReLU}\big(\widehat{r}^s_{i,p,m} - \mathrm{sg}(\widehat{r}^L_{i,p,m})\big)\,, \\ \alpha_{i,p,m} &= \frac{\exp(d_{i,p,m}/\tau_\Delta)}{\sum_{m'} \exp(d_{i,p,m'}/\tau_\Delta)}, \end{aligned} \tag{4}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The late map is recomputed at every forward pass but detached from the auxiliary loss, preventing it from being reshaped merely to enlarge the depth difference.

We compute the differential phrase score as $s^\Delta(I_i, p) = \sum_m \alpha_{i,p,m} r^s_{i,p,m}$. Thus, the normalized depth difference determines where to attend, while the score retains the original shallow cosine similarities. This weighting can emphasize modifier-sensitive regions and shallow pose, contact, or spatial-configuration cues for grounded relations.

### 4.2 Masked Pairwise Sigmoid Objective

Let $\{q_j\}_{j=1}^{N_{\text{comp}}}$ be the compositional phrases in a batch and $b_j$ the owner image of $q_j$. Each image–phrase pair has the auxiliary logit $\ell^\Delta_{i,j} = \frac{s^\Delta(I_i, q_j)}{\tau_{\text{comp}}} + \beta_\Delta$, where $\tau_{\text{comp}}$ is fixed and $\beta_\Delta$ is a training-only learnable bias.

To avoid false negatives from repeated phrases, let $\nu(\cdot)$ denote phrase normalization and $\bar{\mathcal{P}}_i = \{\nu(p) : p \in \mathcal{P}_{\text{comp}}(c_i)\}$. We retain the valid pairs $\mathcal{V} = \big\{(i,j) : i = b_j \vee \nu(q_j) \notin \bar{\mathcal{P}}_i\big\}$. Thus, owner pairs are positive, non-owner exact-text collisions are ignored, and all other valid pairs are negative. With $y^\Delta_{i,j} = \mathbf{1}[i = b_j]$, the masked pairwise sigmoid loss is $\mathcal{L}^\Delta_{\text{comp}} = -\frac{1}{|\mathcal{V}|} \sum_{(i,j)\in\mathcal{V}} \log \sigma\big[(2y^\Delta_{i,j} - 1)\ell^\Delta_{i,j}\big]$.

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \lambda_\Delta \mathcal{L}^\Delta_{\text{comp}}, \tag{5}$$

where $\mathcal{L}_{\text{cap}}$ is the backbone-native global objective. The auxiliary branch and bias are removed at inference. Exact block selection, normalization, collision handling, and bias initialization are detailed in Appendix.

## 5 Experiments

### 5.1 Experimental Setup

**Training data.** We fine-tune all controlled models on a 300K subset of CC3M Sharma et al. (2018), denoted CC300K, using the short captions from DreamLIP Zheng et al. (2024). Qwen3-4B-Instruct-2507 pre-extracts attribute–entity and grounded-relation phrases for the auxiliary loss. We emphasize comparisons under the same CC300K budget and report full-CC3M results for SigLIP and C$^2$LIP only as data-scale references.

| Models | SugarCrepe | | | SugarCrepe++ | | | | ImNet1K | Flickr30k | MSCOCO | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Add | Replace | Swap | Replace | | Swap | | | | | |
| | | | | I2T | TOT | I2T | TOT | | | | |
| OpenAI CLIP-B/32 | 73.0 | 80.0 | 62.7 | 69.5 | 60.5 | 45.7 | 27.4 | 63.3 | 89.0 | 65.4 | 63.7 |
| OpenAI CLIP-B/16 | 72.7 | 80.4 | 62.5 | 70.1 | 60.2 | 43.8 | 23.9 | 68.4 | 90.9 | 67.6 | 64.1 |
| SigLIP ViT-B/16 | 86.5 | 84.1 | 65.8 | 73.8 | 62.8 | 48.0 | 34.7 | **76.1** | 95.2 | 78.9 | 70.6 |
| ***Fine-grained models*** | | | | | | | | | | | |
| FG-CLIP | 84.7 | 85.1 | 69.9 | 75.8 | 67.5 | 51.5 | 38.2 | 69.0 | 95.8 | 78.4 | 71.6 |
| FineCLIP | 85.4 | 85.3 | 66.8 | 72.8 | 68.7 | 43.9 | 26.8 | 55.8 | 93.5 | 79.0 | 67.8 |
| ***Composition-aware models*** | | | | | | | | | | | |
| NegCLIP | 85.8 | 85.0 | **75.3** | 69.1 | 70.9 | 53.4 | 39.1 | 55.7 | 92.4 | 73.9 | 70.1 |
| CE-CLIP | 92.9 | 87.0 | 74.9 | 56.5 | 67.0 | 34.3 | 32.5 | 40.4 | 87.4 | 71.9 | 64.5 |
| CLIC | 89.8 | 85.8 | 72.5 | 76.6 | 58.1 | **59.0** | 27.0 | 66.6 | 91.1 | 67.4 | 69.4 |
| ***Models fine-tuned on CC300k*** | | | | | | | | | | | |
| OpenAI CLIP-B/16 | 86.5 | 84.2 | 68.1 | 73.1 | 72.6 | 45.9 | 34.8 | 65.5 | 95.8 | 76.6 | 70.3 |
| **RELIP (CLIP)** | 89.4 | 85.3 | 69.6 | 74.8 | 73.7 | 47.5 | 37.8 | 66.2 | 96.3 | 78.1 | 71.9 |
| SigLIP ViT-B/16 | 90.0 | 86.1 | 71.5 | 75.0 | 70.3 | 49.8 | 37.9 | 69.5 | 95.6 | 81.7 | 72.7 |
| $C^2$LIP | 91.9 | 87.6 | 73.0 | 77.3 | 75.4 | 51.8 | 43.1 | 69.8 | 96.5 | 81.6 | 74.8 |
| **RELIP (SigLIP)** | **93.9** | **88.8** | 74.5 | **78.7** | **77.4** | 53.9 | **46.3** | 70.5 | **97.4** | **82.3** | **76.4** |
| ***Models fine-tuned by CC3M*** | | | | | | | | | | | |
| SigLIP ViT-B/16 | 87.9 | 85.6 | 69.7 | 73.5 | 67.9 | 49.8 | 36.6 | 75.9 | 95.6 | 80.3 | 75.5 |
| $C^2$LIP | 94.2 | 88.3 | 73.1 | 79.7 | 75.3 | 55.2 | 44.2 | 73.5 | 97.0 | 82.7 | 76.3 |

Table 1: Performance comparison of ViT-B-based models. The CC300K block reports backbone-matched comparisons for CLIP and SigLIP. Full-CC3M results are included as data-scale references.

**Base models and training.** We instantiate RELIP on two 12-block dual encoders: OpenAI CLIP ViT-B/16 and WebLI-pretrained SigLIP ViT-B/16 Zhai et al. (2023). Each backbone retains its native global image–caption objective and uses the same shallow–late differential loss. We fine-tune the full image and text encoders on the identical CC300K subset with $\lambda_\Delta = 0.1$ unless stated otherwise. Optimization details are provided in the Appendix.

**Implementation details.** RELIP uses relative rather than absolute depth. For both backbones, we use blocks $\mathcal{I}_s = \{0, 1, 2, 3\}$ as the shallow stage and block 11 as the late reference. The projected patch tokens and auxiliary branch are used only during training. Fine-tuning RELIP with SigLIP on CC300K takes approximately 30 minutes on three NVIDIA RTX PRO 5000 Blackwell GPUs.

**Evaluation benchmarks.** We evaluate compositional understanding on SugarCrepe Hsieh et al. (2023) and SugarCrepe++ Dumpala et al. (2024), retrieval on Flickr30K Young et al. (2014) and MSCOCO, and zero-shot classification on ImageNet-1K Russakovsky et al. (2015). These evaluations test whether compositional gains preserve general vision–language capabilities.

**Baselines.** We compare with pretrained CLIP and SigLIP, fine-grained models FG-CLIP Xie et al. (2025) and FineCLIP Jing et al. (2024), and composition-aware models NegCLIP Yuksekgonul et al. (2022), CE-CLIP Le Zhang & Agrawal (2023), CLIC Peleg et al. (2025a), and $C^2$LIP Pham et al. (2026). Our controlled comparisons comprise two backbone-matched tracks: CLIP fine-tuning versus RELIP (CLIP), and SigLIP fine-tuning and $C^2$LIP versus RELIP (SigLIP). Within each track, all models use the same CC300K subset and fine-tuning budget.

### 5.2 Main Results

Table 1 compares RELIP with pretrained, fine-grained, and composition-aware vision–language models. Our primary comparisons use backbone-matched CC300K fine-tuning, while full-CC3M results are reported only as data-scale references.

On CLIP ViT-B/16, RELIP improves every reported metric and raises the column-wise average from 70.3 to 71.9, with gains spanning compositional benchmarks, zero-shot classification, and retrieval.

| Models | SugarCrepe | SugarCrepe++ | Flickr30K | MSCOCO |
|---|---|---|---|---|
| SigLIP FT | 83.0 | 60.3 | 95.6 | 81.7 |
| $C^2$LIP | 84.7 | 64.7 | 96.5 | 81.6 |
| $C^2$LIP-Qwen | 84.7 | 64.8 | 96.4 | 81.6 |
| Late-Only | 84.8 | 64.8 | 96.6 | 81.7 |
| Shallow-Only | 85.4 | 65.3 | 96.8 | 82.0 |
| **RELIP** | **86.1** | **66.9** | **97.4** | **82.3** |

Table 2: Controlled ablation using the same CC300K budget. $C^2$LIP-Qwen changes only the phrase source, while all depth variants use the identical Qwen-extracted phrases.

On SigLIP, RELIP reaches $86.1$ on SugarCrepe and $66.9$ on SugarCrepe++, outperforming $C^2$LIP by $1.4$ and $2.2$ points under the same CC300K budget. The improvements span add, replace, and swap subtasks rather than a single perturbation type. RELIP also improves ImageNet-1K, Flickr30K, and MSCOCO to $70.5$, $97.4$, and $82.3$, respectively, raising the column-wise average from $72.7$ to $76.4$. This slightly exceeds the full-CC3M $C^2$LIP reference of $76.3$ while using only the CC300K subset.

## 5.3 Controlled Analysis

### 5.3.1 Phrase Source and Supervision Depth

Unless otherwise noted, analyses in this section use SigLIP ViT-B/16; Table 1 separately evaluates transfer to CLIP.

Table 2 isolates the effects of phrase source, supervision depth, and differential weighting under the same CC300K budget. Replacing $C^2$LIP's original phrases with the Qwen-extracted phrases used by RELIP has negligible effect, with $C^2$LIP-Qwen matching $C^2$LIP on SugarCrepe ($84.7$) and differing by at most $0.1$ points on SugarCrepe++, Flickr30K, and MSCOCO. Thus, phrase source alone does not explain RELIP's gains.

Using the same Qwen phrases, Shallow-Only outperforms Late-Only on SugarCrepe ($85.4$ vs. $84.8$) and SugarCrepe++ ($65.3$ vs. $64.8$), consistent with our depth-wise finding that fine-grained compositional distinctions are better preserved in earlier representations.

RELIP further improves over Shallow-Only to $86.1$ on SugarCrepe and $66.9$ on SugarCrepe++, while also improving retrieval. This isolates the benefit of shallow–late differential weighting beyond shallow supervision alone.

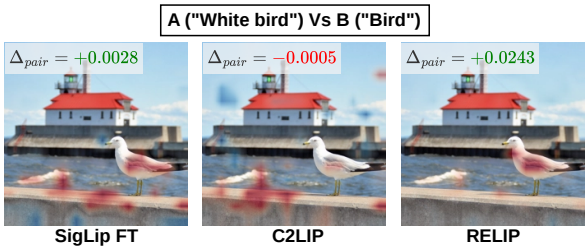
### 5.3.2 Qualitative comparison.

Figure 4 visualizes the post-hoc signed difference $M_{\text{diff}}(m) = d_A(m) - d_B(m)$ between target phrase $A$ and alternative phrase $B$, where warm and cool regions favor $A$ and $B$, respectively. The global margin $\Delta_{\text{pair}} = \cos(g_I, e_A) - \cos(g_I, e_B)$ measures their overall separation: positive values favor the target, values near zero indicate weak discrimination, and negative values favor the alternative.

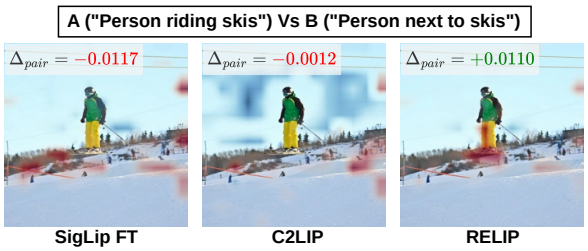
For *white bird* versus *bird*, the margins are $+0.0028$, $-0.0005$, and $+0.0243$ for SigLIP fine-tuning, $C^2$LIP, and RELIP, respectively; RELIP thus shows stronger modifier-specific separation and more target-favoring evidence on the bird region. For *person riding skis* versus *person next to skis*, the corresponding margins are $-0.0117$, $-0.0012$, and $+0.0110$, with RELIP additionally emphasizing the skier and contact region. These examples qualitatively support the controlled results.

### 5.3.3 Hyperparameter and Phrase-Set Robustness

Table 3 examines sensitivity to the compositional-loss weight $\lambda_\Delta$ and phrase-set composition. Compared with the $\lambda_\Delta = 0$ baseline ($83.0/95.6$ on SugarCrepe/Flickr30K), all nonzero settings in $\lambda_\Delta \in [0.05, 0.2]$ consistently improve both metrics, reaching $85.6$–$86.1$ and $96.8$–$97.4$, respectively. Thus, RELIP does not rely on a narrowly tuned loss weight.




(a) Attribute binding




(b) Relation discrimination

Figure 4: Qualitative comparison of post-hoc signed difference maps. Warm regions indicate stronger evidence for target phrase $A$, whereas cool regions indicate stronger evidence for alternative phrase $B$. All maps within each row share the same symmetric color scale centered at zero. $\Delta_{\text{pair}}$ denotes the final global pairwise margin.

| Benchmark | $\lambda_\Delta$ | | | $\mathcal{P}_{\text{comp}}(c_i)$ | | Full |
|---|---|---|---|---|---|---|
| | 0 | 0.05 | 0.2 | w/o Attribute | w/o Relation | |
| SugarCrepe | 83.0 | 85.6 | 86.0 | 85.8 | 85.7 | **86.1** |
| Flickr30k | 95.6 | 96.8 | 97.0 | 96.8 | 96.9 | **97.4** |

Table 3: Robustness to $\lambda_\Delta$ and the compositional phrase set. Full uses $\lambda_\Delta = 0.1$ with both phrase categories.

With $\lambda_\Delta = 0.1$, removing either attribute–entity or grounded-relation phrases retains most of the improvement, while using both gives the strongest overall result. The two phrase categories therefore provide complementary rather than individually indispensable supervision. Detailed subtask results are provided in the Appendix.

### 5.3.4 Efficiency

RELIP adds no inference-time modules or parameters. On SigLIP ViT-B/16, it requires 10.3 GiB and 1.46 seconds per optimizer step, corresponding to 24% more memory and 26% more step time than standard fine-tuning. It remains lighter than $C^2$LIP, using 27% less memory and 72% less step time under identical hardware. Full profiling settings are provided in the Appendix.

## 6 Conclusion

We identify complementary depth specialization in contrastive vision–language encoders: noun-phrase selectivity strengthens toward later layers, whereas modifier- and relation-sensitive distinctions attenuate. Motivated by this finding, RELIP uses shallow compositional evidence with a detached late semantic reference while preserving the backbone's native objective and inference architecture. Controlled experiments show that its gains arise beyond phrase source or shallow supervision alone, and RELIP consistently improves both CLIP and SigLIP without sacrificing general vision–language performance. These results highlight encoder depth as an important dimension of compositional adaptation.

## REFERENCES


Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. ColorSwap: A color and word order dataset for multimodal evaluation. In *Findings of the Association for Computational Linguistics: ACL*, pp. 1716–1726, 2024.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150, 2023a.

Sivan Doveh, Assaf Arbelle, Leonid Karlinsky, Rameswar Panda, Rogerio Feris, Aude Oliva, and Giridharan Raja. Teaching structured vision & language concepts to vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

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.

B Gurung, David T Hoffmann, and Thomas Brox. Common data properties limit object-attribute binding in clip. In *German Conference on Pattern Recognition (GCPR)*, 2025.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Ranjan. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3221–3239, 2023.

Bora Kargi, Arnas Uselis, and Seong Joon Oh. Half-truths break similarity-based retrieval. *arXiv preprint arXiv:2602.23906*, 2026.

et al. Kwon. Enhancing compositional reasoning in CLIP via reconstruction and alignment of text descriptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Rabiul Awal Le Zhang and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding. *arXiv preprint arXiv:2306.08832*, 3, 2023.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen H. Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL*, pp. 1487–1500, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. CLIPS: An enhanced CLIP framework for learning with synthetic captions. *arXiv preprint arXiv:2411.16828*, 2024.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3170–3180, 2023.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8253–8280, 2022.

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. *Advances in neural information processing systems*, 37:32731–32760, 2024.

Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in clip with efficient fine-tuning. *arXiv preprint arXiv:2505.24424*, 2025a.

Amit Peleg et al. Advancing compositional awareness in CLIP with efficient fine-tuning. *arXiv preprint*, 2025b.

Hai Xuan Pham, David Hoffmann, Ricardo Guerrero, and Brais Martinez. No hard negatives required: Concept centric learning leads to compositionality without degrading zero-shot capabilities of contrastive models. In *CVPR*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Ankit Vani, Bac Nguyen, Samuel Lavoie, Ranjay Krishna, and Aaron Courville. Sparo: Selective attention for robust and compositional transformer encodings for vision. In *European Conference on Computer Vision*, pp. 233–251. Springer, 2024.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: Vlm with fine-grained language-informed image representations. *arXiv preprint arXiv:2412.03561*, 2024.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

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.

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? *arXiv preprint arXiv:2210.01936*, 2022.

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 *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13774–13784, 2024.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. 2024.

## A Additional Implementation Details

**Layer selection and projected patch tokens.** For both 12-block backbones, OpenAI CLIP ViT-B/16 and SigLIP ViT-B/16, we use the first four blocks $\mathcal{I}_s = \{0, 1, 2, 3\}$ as the shallow stage and block $L = 11$ as the late reference. At each selected block, we extract the pre-attention normalized tokens from the $\mathrm{LN}_1$ output and discard the class token when present. We apply the backbone-specific visual projection to tokens from all selected depths and L2-normalize the resulting patch features before computing patch–phrase cosine similarities. Phrase embeddings are obtained after the corresponding text projection and L2-normalized in the same joint embedding space. Each backbone retains its native global readout and objective; the projected intermediate patch tokens are used only by the auxiliary branch.

**Patch-wise normalization and differential weighting.** For a relevance map $r \in \mathbb{R}^M$, we define patch-wise z-score normalization as

$$\widehat{r}_m = \frac{r_m - \mu(r)}{\mathrm{Std}(r) + \epsilon}, \tag{6}$$

where the statistics are computed over the $M$ spatial patches. This normalization removes depth-dependent scale and offset differences before comparing shallow and late relevance patterns. The differential evidence is

$$d_{i,p,m} = \mathrm{ReLU}\big(\widehat{r}^{s}_{i,p,m} - \mathrm{sg}(\widehat{r}^{L}_{i,p,m})\big)\,. \tag{7}$$

The late map is recomputed at every forward pass but receives no gradient from $\mathcal{L}^{\Delta}_{\mathrm{comp}}$; the late visual blocks remain trainable through the backbone-native global objective, denoted $\mathcal{L}_{\mathrm{cap}}$. The subsequent spatial softmax assigns larger weights to positive shallow–late residuals without hard-masking the remaining patches. Normalized maps are used only to derive the spatial weights, whereas the final phrase score retains the original shallow cosine similarities.

**Phrase normalization and collision handling.** Phrase normalization $\nu(\cdot)$ is used only to detect repeated phrases and does not modify the text passed to the text encoder. We lowercase each phrase, remove punctuation and English articles (*a*, *an*, and *the*), and collapse consecutive whitespace before exact-string matching. For each image $i$, we define

$$\bar{\mathcal{P}}_i = \{\nu(p) \mid p \in \mathcal{P}_{\mathrm{comp}}(c_i)\}\,. \tag{8}$$

For phrase $q_j$ owned by image $b_j$, the owner pair is positive, non-owner images containing the same normalized phrase are ignored, and all remaining valid image–phrase pairs are negative. Ignored pairs contribute to neither the numerator nor the denominator of the masked sigmoid loss. Phrase sets and owner indices are gathered across data-parallel workers before mask construction, ensuring identical handling of within-worker and cross-worker collisions.

**Auxiliary-logit calibration.** The compositional branch uses a fixed temperature $\tau_{\mathrm{comp}}$ and a separate training-only learnable scalar bias:

$$\ell^{\Delta}_{i,j} = \frac{s^{\Delta}(I_i, q_j)}{\tau_{\mathrm{comp}}} + \beta_{\Delta}. \tag{9}$$

We initialize $\beta_{\Delta}$ to zero and optimize it jointly with the model. The auxiliary branch uses separate calibration because global image–caption similarities and differentially pooled patch–phrase scores have different distributions. The bias is discarded together with the auxiliary branch at inference.

## B Training and Optimization Details

Unless otherwise noted, we use the same training configuration for OpenAI CLIP ViT-B/16 and SigLIP ViT-B/16.

**Training Hyperparameters** Table S1 summarizes the hyperparameters used in the controlled CC300K experiments. The auxiliary scalar bias $\beta_{\Delta}$ is initialized to zero and jointly optimized with the model.

| Setting | Value |
|---|---|
| Z-score stability constant $\epsilon$ | $10^{-6}$ |
| L2-normalization constant | $10^{-8}$ |
| Spatial temperature $\tau_{\Delta}$ | 0.07 |
| Image–phrase temperature $\tau_{\text{comp}}$ | 0.07 |
| Compositional-loss weight $\lambda_{\Delta}$ | 0.1 |
| Initial auxiliary bias $\beta_{\Delta}$ | 0 |
| Optimizer | AdamW |
| AdamW $(\beta_1, \beta_2)$ | $(0.9, 0.98)$ |
| AdamW $\epsilon$ | $10^{-6}$ |
| Learning rate | $10^{-5}$ |
| Weight decay | 0.1 |
| Training epochs | 5 |
| Warmup | 1 epoch |
| Learning-rate schedule | Linear warmup, cosine decay |
| Mixed precision | FP16 AMP |
| Image resolution | $224 \times 224$ |
| Gradient clipping | 1.0 |

Table S1: Training hyperparameters used for the controlled CC300K experiments. Weight decay is not applied to biases, one-dimensional parameters, or logit-calibration parameters.

**Batching and Randomness** Training is distributed over three GPUs with a per-GPU micro-batch size of 32, yielding a global per-forward batch size of 96. We accumulate gradients over eight micro-steps, resulting in an effective batch size of 768. We use base seed 42 and rank-specific seeds $42 + r$ for data-parallel rank $r$. Each reported controlled configuration is obtained from one distributed training run.

**Computing Environment** Experiments are conducted on Ubuntu 22.04 LTS using Python 3.10.19, PyTorch 2.10.0 with CUDA 12.8, and cuDNN 9.10.2. Training uses three NVIDIA RTX PRO 5000 Blackwell GPUs, each with 48,935 MiB of memory. Our implementation is based on a local OpenCLIP 4.0.0 development build (commit `fe3631f`), together with torchvision 0.25.0, timm 1.0.28, and Transformers 5.2.0.

## C Detailed Benchmark Results

This section reports the complete SugarCrepe and SugarCrepe++ subtask breakdowns. Each perturbation-family average is the unweighted mean of its displayed subtask accuracies, whereas the final *Average* column reports accuracy over all benchmark examples.

**SugarCrepe Results** Table S2 shows that RELIP improves compositional performance consistently across both backbones and all SugarCrepe subtasks. On CLIP ViT-B/16, RELIP raises the Replace, Swap, and Add averages from 84.2, 68.1, and 86.6 to 85.3, 69.6, and 89.4, respectively, increasing the overall score from 80.3 to 82.0.

The gains are larger on SigLIP. Relative to standard SigLIP fine-tuning, RELIP improves the Replace, Swap, and Add averages by 2.7, 3.0, and 3.8 points, yielding an overall improvement from 83.0 to 86.1. RELIP also outperforms $C^2$LIP on every reported subtask, improving the three family averages from 87.6/73.0/91.9 to 88.8/74.5/93.8. These results show that the overall gain is distributed across object-, attribute-, and relation-sensitive perturbations rather than being driven by a single subset.

**SugarCrepe++ Results** Table S3 provides the full SugarCrepe++ breakdown. On CLIP ViT-B/16, RELIP improves every object, attribute, and relation subtask, raising the Replace-I2T, Replace-TOT, Swap-I2T, and Swap-TOT averages from 73.1, 72.6, 45.9, and 34.8 to 74.8, 73.7, 47.5, and 37.9, respectively. The overall score consequently increases from 59.9 to 61.6.

On SigLIP, RELIP improves the corresponding averages over standard fine-tuning by 3.7, 7.1, 4.1, and 8.4 points, reaching an overall score of 66.9. Compared with $C^2$LIP, RELIP improves every

| Models | Replace | | | | Swap | | | Add | | | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Obj | Attr | Rel | Avg | Obj | Attr | Avg | Obj | Attr | Avg | |
| SigLIP ViT-B/16 | 95.3 | 86.7 | 70.3 | 84.1 | 60.0 | 71.5 | 65.8 | 89.1 | 83.8 | 86.5 | 79.5 |
| OpenAI CLIP-B/32 | 90.9 | 80.0 | 69.2 | 80.0 | 61.2 | 64.1 | 62.7 | 77.2 | 68.8 | 73.0 | 73.1 |
| OpenAI CLIP-B/16 | 93.5 | 81.1 | 66.7 | 80.4 | 60.0 | 65.0 | 62.5 | 78.5 | 66.9 | 72.7 | 73.1 |
| FG-CLIP | 95.9 | 87.1 | 72.2 | 85.1 | 66.5 | 73.3 | 69.9 | 87.6 | 81.8 | 84.7 | 80.6 |
| FineCLIP | 95.6 | 85.2 | 75.0 | 85.3 | 63.3 | 70.3 | 66.8 | 90.0 | 80.8 | 85.4 | 80.0 |
| NegCLIP | 92.7 | 85.9 | 76.5 | 85.0 | **75.2** | 75.4 | **75.3** | 88.8 | 82.8 | 85.8 | 82.5 |
| CE-CLIP | 93.1 | 88.8 | 79.0 | 87.0 | 72.8 | 77.0 | 74.9 | 92.4 | 93.4 | 92.9 | 85.2 |
| CLIC | 95.6 | 86.6 | 75.3 | 85.8 | 71.0 | 74.0 | 72.5 | 88.4 | 91.2 | 89.8 | 83.1 |
| OpenAI CLIP-B/16 | 95.3 | 85.8 | 71.5 | 84.2 | 67.8 | 68.5 | 68.1 | 86.9 | 86.3 | 86.6 | 80.3 |
| **RELIP (CLIP)** | 96.7 | 87.1 | 72.1 | 85.3 | 69.9 | 69.3 | 67.6 | 89.6 | 89.1 | 89.4 | 82.0 |
| SigLIP ViT-B/16 | 96.4 | 88.2 | 73.8 | 86.1 | 65.6 | 77.3 | 71.5 | 89.5 | 90.5 | 90.0 | 83.0 |
| $C^2$LIP | 96.4 | 89.6 | 77.0 | 87.6 | 68.9 | 76.9 | 73.0 | 91.1 | 92.8 | 91.9 | 84.7 |
| **RELIP (SigLIP)** | **96.8** | **90.0** | **79.6** | **88.8** | 70.1 | **78.8** | 74.5 | **93.3** | **94.4** | **93.8** | **86.1** |

Table S2: Full SugarCrepe subtask breakdown. The bottom block reports controlled CC300K fine-tuning results. Bold and underlined values denote the best and second-best results, respectively.

| Models | Replace - I2T | | | | Replace - TOT | | | | Swap - I2T | | | Swap - TOT | | | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Obj | Attr | Rel | Avg | Obj | Attr | Rel | Avg | Obj | Attr | Avg | Obj | Attr | Avg | |
| SigLIP ViT-B/16 | 91.2 | 75.5 | 54.8 | 73.8 | 79.2 | 64.2 | 45.0 | 62.8 | 39.6 | 56.3 | 48.0 | 22.9 | 46.4 | 34.7 | 57.5 |
| OpenAI CLIP-B/32 | 86.7 | 65.6 | 56.3 | 69.5 | 83.7 | 59.3 | 38.6 | 60.5 | 46.1 | 45.2 | 45.7 | 19.1 | 35.6 | 27.4 | 53.6 |
| OpenAI CLIP-B/16 | 89.6 | 67.6 | 53.2 | 70.1 | 84.4 | 57.2 | 39.0 | 60.2 | 39.2 | 48.4 | 43.8 | 16.3 | 31.4 | 23.9 | 52.6 |
| FG-CLIP | 92.6 | 76.8 | 58.0 | 75.8 | 90.6 | 67.8 | 44.2 | 67.5 | 47.8 | 55.3 | 51.5 | 29.4 | 47.0 | 38.2 | 60.9 |
| FineCLIP | 90.8 | 70.7 | 57.0 | 72.8 | 91.3 | 67.9 | 47.0 | 68.7 | 39.6 | 48.2 | 43.9 | 20.8 | 32.7 | 26.8 | 56.6 |
| CE-CLIP | 71.9 | 52.0 | 45.5 | 56.5 | 86.3 | 64.2 | 50.5 | 67.0 | 36.3 | 32.3 | 34.3 | 28.6 | 36.3 | 32.5 | 50.4 |
| NegCLIP | 87.0 | 67.1 | 53.1 | 69.1 | **93.3** | 70.7 | 48.6 | 70.9 | 51.8 | 55.0 | 53.4 | 27.8 | 50.3 | 39.1 | 60.5 |
| CLIC | 91.6 | 75.9 | 62.3 | 76.6 | 84.7 | 52.4 | 37.3 | 58.1 | **55.9** | 62.2 | **59.0** | 22.9 | 31.2 | 27.0 | 57.6 |
| OpenAI CLIP-B/16 | 91.1 | 71.8 | 56.5 | 73.1 | 92.0 | 76.0 | 49.8 | 72.6 | 42.3 | 49.6 | 45.9 | 42.2 | 27.4 | 34.8 | 59.9 |
| **RELIP (CLIP)** | 92.4 | 74.2 | 57.9 | 74.8 | 93.1 | 76.8 | 51.3 | 73.7 | 44.3 | 50.6 | 47.5 | 46.0 | 29.7 | 37.9 | 61.6 |
| SigLIP ViT-B/16 | 91.4 | 76.7 | 56.9 | 75.0 | 90.4 | 72.7 | 47.7 | 70.3 | 42.9 | 56.8 | 49.8 | 23.1 | 52.8 | 37.9 | 60.3 |
| $C^2$LIP | 93.2 | 76.6 | 62.2 | 77.3 | 89.8 | 78.7 | 57.7 | 75.4 | 42.0 | 61.6 | 51.8 | 29.0 | 57.2 | 43.1 | 64.7 |
| **RELIP (SigLIP)** | **93.6** | **78.5** | **64.0** | **78.7** | 92.2 | **82.3** | **57.8** | **77.4** | 44.1 | **63.8** | 53.9 | **31.4** | **61.1** | **46.3** | **66.9** |

Table S3: Full SugarCrepe++ subtask breakdown. The bottom block reports controlled CC300K fine-tuning results. Bold and underlined values denote the best and second-best results, respectively.

reported subtask, raising the four family averages by 1.4, 2.0, 2.1, and 3.2 points. The consistent gains across replace and swap settings indicate that the improvement is not tied to a single corruption type.

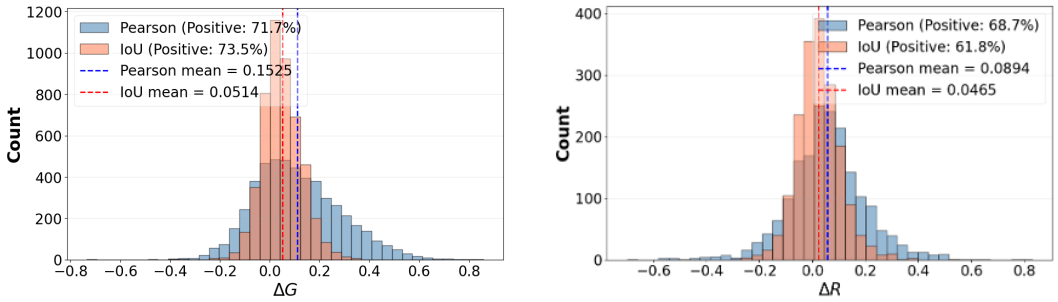
## D Additional Ablations

This section provides the full subtask breakdowns for the loss-weight and phrase-set ablations summarized in the Hyperparameter and Phrase-Set Robustness section. Unless otherwise noted, all experiments use SigLIP ViT-B/16 and the same CC300K fine-tuning setting as the main experiments. Following $C^2$LIP, all Flickr30K and MSCOCO retrieval results are reported using Recall@5.

**Loss-Weight Sensitivity** Table S4 reports the detailed results for $\lambda_\Delta \in \{0, 0.05, 0.1, 0.2\}$. Removing the compositional objective ($\lambda_\Delta = 0$) yields 83.0 on SugarCrepe and 95.6 on Flickr30K. Introducing the differential loss consistently improves both metrics: $\lambda_\Delta = 0.05$ reaches $85.6/96.8$, while $\lambda_\Delta = 0.2$ reaches $86.0/97.0$ on SugarCrepe/Flickr30K. The default setting $\lambda_\Delta = 0.1$ achieves the strongest joint result, reaching 86.1 and 97.4, respectively. The detailed subtask results show that

| $\lambda_\Delta$ | SugarCrepe | | | | | | | | Flickr30k |
|---|---|---|---|---|---|---|---|---|---|
| | Replace | | | Swap | | Add | | **Average** | |
| | Object | Attribute | Relation | Object | Attribute | Object | Attribute | | |
| 0 | 96.3 | 88.2 | 73.8 | 65.5 | 77.3 | 89.4 | 90.5 | 83.0 | 95.6 |
| 0.05 | 97.0 | 89.7 | 78.9 | 68.3 | 78.5 | 93.1 | **94.4** | 85.6 | 96.8 |
| 0.1 | 96.8 (+0.5) | **90.0** (+1.8) | 79.6 (+5.8) | **70.1** (+4.6) | 78.8 (+1.5) | **93.3** (+3.9) | **94.4** (+3.9) | **86.1** (+3.1) | **97.4** (+5.2) |
| 0.2 | **97.1** (+0.8) | 89.5 (+1.3) | **79.7** (+5.9) | 69.5 (+4.0) | **79.0** (+1.7) | 93.1 (+3.7) | **94.4** (+3.9) | 86.0 (+3.0) | 97.0 (+4.8) |

Table S4: Sensitivity to the compositional-loss weight $\lambda_\Delta$ under the controlled CC300K setting.




(a) Noun-phrase gap differences

(b) Modifier-margin differences

Figure S1: **Shallow–late endpoint distributions in CLIP.** (a) Sample-level noun-phrase gap differences $\Delta G = G_{\text{late}} - G_{\text{shallow}}$. (b) Phrase-pair-level modifier-margin differences $\Delta R = B_{\text{shallow}} - B_{\text{late}}$. Positive values indicate stronger late noun-phrase selectivity in (a) and stronger shallow modifier-sensitive specificity in (b). Dashed lines denote the mean differences, and the legends report the proportion of positive observations.

the improvement is not restricted to a single perturbation type, and performance remains stable across the tested nonzero weights.

| Attribute Entity | Grounded Relation | SugarCrepe | | | | | | | | Flickr30K |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Replace | | | Swap | | Add | | **Average** | |
| | | Object | Attribute | Relation | Object | Attribute | Object | Attribute | | |
| ✓ | | 96.8 | 89.6 | 79.0 | 69.4 | 78.2 | 93.4 | 94.3 | 85.8 | 96.8 |
| | ✓ | 96.5 | 89.7 | 79.0 | 69.1 | 78.5 | 93.0 | 94.1 | 85.7 | 96.9 |

Table S5: Effect of the compositional phrase categories with $\lambda_\Delta = 0.1$.

**Phrase-Set Composition** Table S5 examines the contribution of the two phrase categories while fixing $\lambda_\Delta = 0.1$. Removing attribute–entity phrases yields 85.8 on SugarCrepe and 96.8 on Flickr30K, whereas removing grounded-relation phrases yields 85.7 and 96.9, respectively. Both variants retain most of the improvement over caption-only fine-tuning, indicating that either phrase category provides useful compositional supervision. Nevertheless, using both categories produces the best overall performance, with 86.1 on SugarCrepe and 97.4 on Flickr30K. These results suggest that attribute–entity and grounded-relation phrases provide complementary rather than individually indispensable supervision.

## E Extended Depth Analysis

**Endpoint Distributions in CLIP** Figure S1 visualizes the shallow–late endpoint differences underlying the sample-level trends reported in the main text. For noun-phrase selectivity, we define

$$\Delta G = G_{\text{late}} - G_{\text{shallow}},$$

where $G_{\text{shallow}}$ and $G_{\text{late}}$ denote the stage-averaged phrase gaps defined in empirical motivation section of the main paper. Positive values therefore indicate stronger noun-phrase selectivity at late depth. The difference is positive for $71.7\%$ of samples under Pearson correlation and $73.5\%$ under top-$q$ IoU.

For modifier-sensitive binding, we define

$$\Delta R = B_{\text{shallow}} - B_{\text{late}},$$

such that positive values indicate stronger modifier-induced specificity at shallow depth. The difference is positive for $68.7\%$ of phrase pairs under Pearson correlation and $61.8\%$ under top-$q$ IoU. The histogram is computed at the phrase-pair level, whereas the Wilcoxon tests reported in the main text first average multiple phrase pairs within each image and then perform paired testing over images.

**Representative Layer-Wise Trajectories in CLIP** Figure S2a provides representative trajectories across all 24 visual blocks of CLIP ViT-L/14. The noun-phrase examples show that pairwise phrase gaps remain relatively small in shallow layers and increase toward later layers under both Pearson correlation and top-$q$ IoU. In contrast, the same-head modifier examples exhibit their largest margins in earlier layers and generally weaken as depth increases, sometimes reaching near-zero or negative values in the late stage.

These examples illustrate that the stage-level trends in Figure 2 of the main paper are not produced solely by comparing two selected endpoints. Intermediate blocks typically trace a gradual transition between shallow and late representations. Because the examples are illustrative rather than aggregate statistics, the dataset-level evidence remains the stage and distribution analyses reported in the main text and above.

**Cross-Backbone Representative Trajectories** Figure S2 compares representative full-depth trajectories from CLIP ViT-L/14 and SigLIP ViT-B/16. Across both backbones, noun-phrase gaps generally increase toward later layers, whereas same-head modifier margins weaken with depth under both Pearson correlation and top-$q$ IoU. The intermediate blocks trace a gradual transition, suggesting that the specialization follows relative encoder depth rather than a particular endpoint or absolute layer index.

## F Depth Specialization on SigLIP

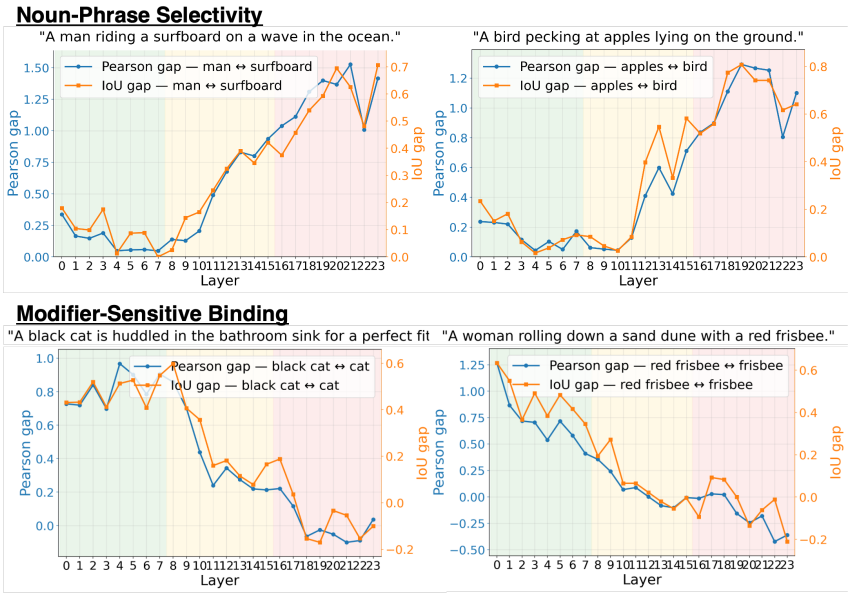
To test whether the observed specialization is specific to the CLIP probe backbone, we repeat the three depth-wise analyses on the WebLI-pretrained SigLIP ViT-B/16. We use the same MSCOCO image–caption pairs, phrase construction, and probe definitions as in empirical motivation section of the main paper. The 12 visual blocks are grouped into shallow (0–3), middle (4–7), and late (8–11) stages. For all three probes, the Pearson-based shallow–late differences are significant ($p < .001$).

**Noun-phrase selectivity.** Figure S3a shows that noun-phrase selectivity again strengthens with depth. The mean phrase gap increases from $0.1250$ in shallow representations to $0.1387$ in middle and $0.1621$ in late representations under Pearson correlation. Top-$q$ IoU exhibits the same ordering, increasing from $0.0828$ to $0.0918$ and $0.1096$. At the sample level, the shallow–late difference $\Delta G = G_{\text{late}} - G_{\text{shallow}}$ is positive for $64.3\%$ of samples under Pearson and $59.0\%$ under top-$q$ IoU, with mean differences of $0.0371$ and $0.0267$, respectively. Thus, as in CLIP, caption-conditioned relevance becomes increasingly uneven across noun phrases toward later SigLIP layers.

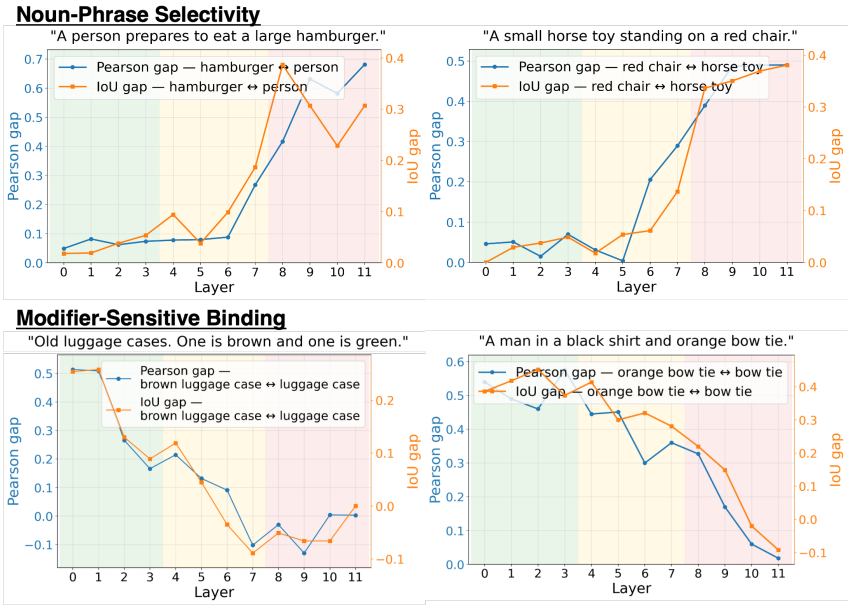
**Modifier-sensitive binding.** The complementary trend is observed for same-head modifier distinctions (Figure S3b). The mean binding margin decreases from $0.1338$ to $0.1092$ and $0.0992$ under Pearson correlation, and from $0.0816$ to $0.0603$ and $0.0537$ under top-$q$ IoU. The endpoint difference $\Delta R = B_{\text{shallow}} - B_{\text{late}}$ is positive for $64.5\%$ and $62.7\%$ of phrase pairs, with mean differences of $0.0342$ and $0.0276$, respectively. These results indicate that the additional spatial specificity introduced by a modifier progressively attenuates across SigLIP depth.

**Relation-sensitive discrimination.** Figure S4b shows the same attenuation for grounded relations. All five relation-family means follow

$$C_{\text{shallow}} > C_{\text{middle}} > C_{\text{late}}.$$




(a) CLIP ViT-L/14

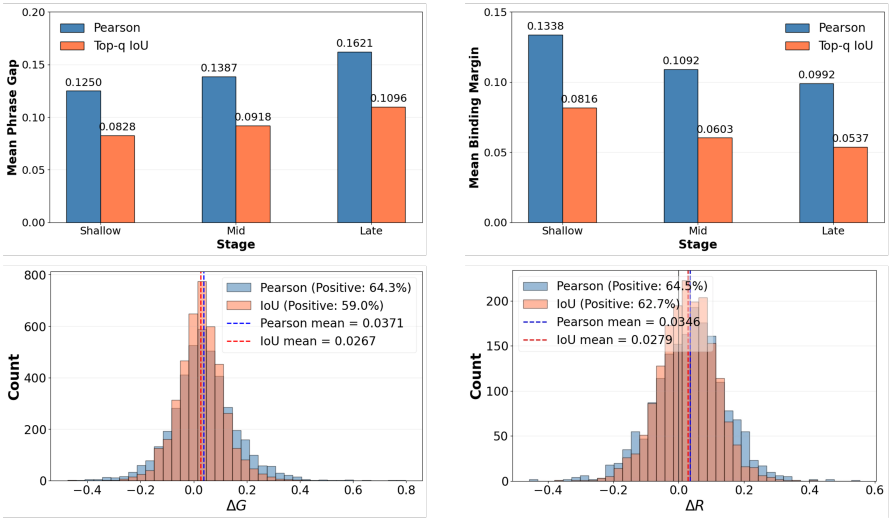



(b) SigLIP ViT-B/16

Figure S2: **Representative full-depth trajectories in CLIP and SigLIP.** For each backbone, noun-phrase examples show gaps that generally increase toward later layers, whereas same-head modifier examples show margins that weaken with depth. Blue and orange curves denote Pearson correlation and top-$q$ IoU, respectively; shaded regions denote the backbone-specific shallow, middle, and late stages.

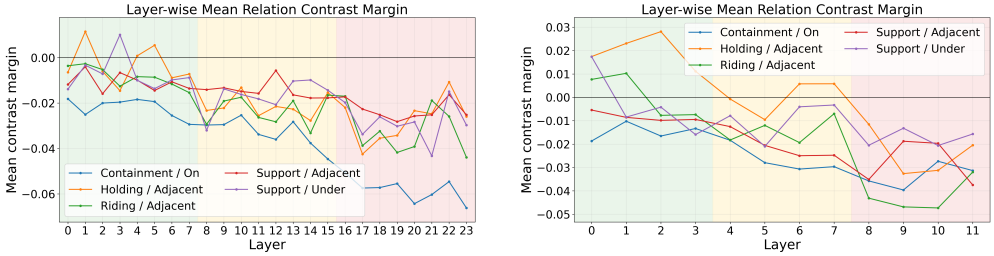
Several target relations that remain positively separated in shallow representations, including holding–adjacency, riding–adjacency, and support–under, become negative toward the final stage. The consistent decline across families indicates that target relations become less separable from their closest alternatives as visual depth increases.

Taken together, SigLIP reproduces the complementary specialization observed in CLIP: later representations exhibit stronger broad noun-phrase selectivity, whereas shallow representations retain




(a) Noun-phrase selectivity (b) Modifier-sensitive binding

Figure S3: Depth-wise specialization in SigLIP. (a) Mean noun-phrase gaps across depth and sample-level endpoint differences $\Delta G = G_{\text{late}} - G_{\text{shallow}}$. (b) Mean modifier-binding margins and endpoint differences $\Delta R = B_{\text{shallow}} - B_{\text{late}}$. Positive $\Delta G$ indicates stronger late noun-phrase selectivity, whereas positive $\Delta R$ indicates stronger shallow modifier-sensitive specificity.




(a) CLIP ViT-L/14 (b) SigLIP ViT-B/16

Figure S4: Family-wise layer trajectories for grounded-relation discrimination. Each curve shows the mean contrast margin for one relation-family comparison. Across both CLIP and SigLIP, contrast margins tend to become smaller (more negative) toward later layers, indicating weaker separation from the hardest confusable alternative.

larger modifier- and relation-sensitive margins. This provides representation-level evidence that the depth-asymmetric principle behind RELIP generalizes across both backbones.

## AI USE STATEMENT

We used generative AI assistants, including ChatGPT, during research ideation and manuscript preparation. Specifically, they were used to discuss and refine hypotheses, provide feedback on experimental design and methodological framing, summarize and compare related literature, assist in interpreting experimental results, and draft or edit portions of the manuscript for clarity and concision.

Separately, Qwen3-4B-Instruct-2507 was used as part of the reported experimental pipeline to identify entity, attribute–entity, and grounded-relation spans already present in image captions, as described in the main text. It was not used to regenerate captions or create benchmark labels.

All AI-assisted suggestions were reviewed by the authors. Experimental claims were checked against the corresponding results, and literature and citation-related claims were manually verified. The authors take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

## Ethics statement

This work does not involve new human-subject studies or the collection of new personal data. We use existing vision–language datasets, benchmarks, and pretrained models for research purposes. As with other vision–language models trained or evaluated on web-derived data, the models studied here may inherit biases and limitations present in their pretraining and evaluation data. RELIP is designed to improve compositional representation and does not explicitly address fairness, bias, privacy, or safety; improved compositional performance should therefore not be interpreted as improvement along these dimensions.

## Reproducibility statement

We provide the methodological and experimental details needed to reproduce our results throughout the main paper and appendix. The main text specifies the depth-wise probing procedure, RELIP objective, model backbones, training data, and evaluation protocol. Appendix A details layer selection, patch-token projection and normalization, differential weighting, phrase-collision handling, and auxiliary-logit calibration. Appendix B reports the complete training hyperparameters, batching and randomness settings, and computing environment. Appendices C–D provide full benchmark and ablation results, while Appendices E–F provide additional depth-wise analyses and the corresponding SigLIP study.