FROM CONTROLLED EDITS TO COMPLEX BINDINGS: EVALUATING AND IMPROVING STRUCTURE-AWARE VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP is effective at global image–text matching, and Structure-CLIP further improves discrimination on template-based attribute and relation perturbations. However, whether these gains extend to descriptions containing multiple entities, modifiers, and relations remains unexplored. A structural error in such descriptions may concern not only attribute or relation direction but also host assignment or spatial layout. We construct 666 image–text instances over six structural categories. Each instance pairs a natural image with a complex multi-entity caption P_1, a meaning-preserving paraphrase P_2, and a negative N that changes one visually verifiable binding. This $P_1/P_2/N$ design evaluates whether a structural decision is stable across two valid expressions and also supplies two views of the same binding during training. An exploratory evaluation of pretrained checkpoints shows uneven transfer. Structure-CLIP remains strong on relation direction, but improves little or degrades on host, left–right, and depth bindings. We therefore introduce Complex-SCLIP, which applies the existing knowledge-enhanced encoder to both positives and jointly optimizes dual ranking, multi-positive contrastive learning, and positive consistency against the same structural negative. On the 240-instance test split, Complex-SCLIP reaches 69.2% Both-accuracy, outperforming Structure-CLIP (41.7%), five published compositional methods (32.1–39.6%), and an in-domain Structure-CLIP fine-tuning baseline trained with the released objective (48.3%). The improvement is consistent across three seeds and transfers to SUGARCREPE++. COCO 5K Mean Recall remains above CLIP (61.7% vs. 60.3%), though below the Structure-CLIP starting point.

1 INTRODUCTION

Contrastive vision–language models such as CLIP (Radford et al., 2021) learn strong image–text matching from large-scale data, yet often fail to capture how words are bound into structure (Yüksekgönül et al., 2023; Thrush et al., 2022). An image of an *astronaut riding a horse* and the caption *a horse riding an astronaut* share almost every token, although the subject–object assignment is reversed. CLIP-style models can assign similar scores to these captions despite the reversed roles. Structure-CLIP (Huang et al., 2024) addresses part of this problem with scene-graph knowledge and deterministic structural negatives, and reports clear gains on template-based attribute and relation tests.

Real-world captions typically describe several objects, attributes, and interactions in the same sentence, rather than isolating a single entity or relation. Because most of the caption remains correct, a model may match the image using unaffected content and ignore the edited binding (Hsieh et al., 2023). Such descriptions also involve perturbations beyond the original attribute and relation templates, such as assigning clothing or a carried object to the wrong host, or reversing the left–right, above–below, or front–behind order of two visible entities (Liu et al., 2023; Kamath et al., 2023). It remains unclear whether the gains reported for Structure-CLIP are reliable when a single structural error is embedded in a description with multiple entities, modifiers, and relations.

We study this transfer with a controlled $P_1/P_2/N$ task. Each description is a complex multi-entity sentence containing several modifiers and relations. For the same image, P_1 is an original correct

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.

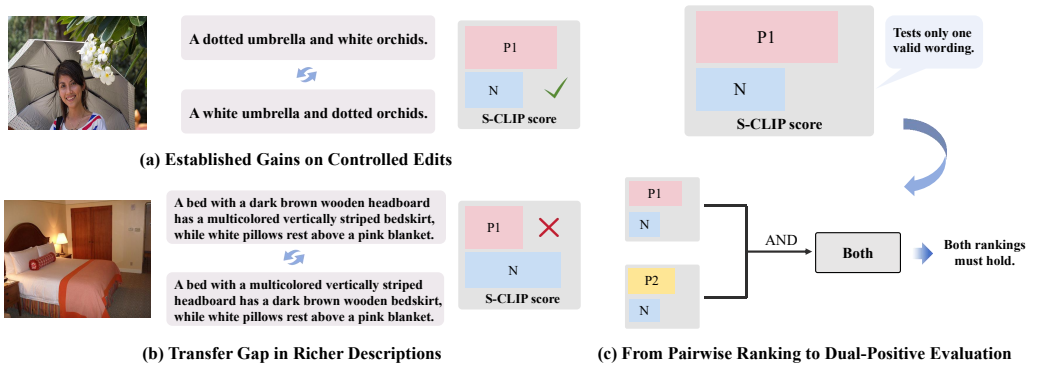

Figure 1: Motivation. (a) Structure-CLIP improves pairwise ranking on controlled structural edits. (b) When the edited binding is embedded in a description containing several other facts, the same pairwise decision can fail. (c) Adding a meaning-preserving P_2 extends the evaluation from one valid wording to the joint Both criterion. Similarity bars schematically indicate only the ranking outcome.

description, P_2 is a meaning-preserving paraphrase, and N changes exactly one target binding while preserving the remaining visual facts. The inclusion of P_2 is motivated by recent work showing that hard-negative fine-tuning can degrade valid-paraphrase acceptance (Dumpala et al., 2024; Kamath et al., 2024). P_2 serves both evaluation and training: requiring $P_1 > N$ and $P_2 > N$ tests whether the structural decision survives a valid rephrasing, while during training P_2 supplies a second linguistic view of the same visual binding, discouraging the shared text encoder from fitting one caption form.

As shown in Figure 1, the evaluation extends from pairwise ranking to a stricter dual-positive criterion.

We construct 666 instances across attribute, relation, host, and three spatial categories. Evaluating pretrained CLIP and Structure-CLIP checkpoints on the complete collection shows that performance varies across categories. Specifically, Structure-CLIP improves P_1-accuracy strongly for relation direction (+30.4%) but provides limited benefit for host (+4.4%) and no gain for left–right (-6.1%) or depth (-1.7%). We introduce Complex-SCLIP. Starting from the Structure-CLIP checkpoint, the method applies the same knowledge-enhanced encoder separately to P_1 and P_2, ranks both against a shared structural negative, and aligns the two valid descriptions through a batch-level multi-positive objective and consistency constraints.

We compare Complex-SCLIP with CLIP, Structure-CLIP, five published compositional checkpoints (including CLIC (Peleg et al., 2025), a multi-positive method), and a Structure-CLIP P1/N continual-fine-tuning baseline using the released objective. Ablations identify explicit P_2 supervision as the largest source of improvement, followed by multi-positive learning and positive-side knowledge enhancement. We also evaluate SUGARCREPE++ transfer and COCO 5K retrieval preservation.

Our contributions can be summarized as follows:

- a controlled $P_1/P_2/N$ resource of 666 image–text instances covering six structural categories in descriptions with multiple entities, modifiers, and relations, with one visually verifiable structural edit per instance;
- an empirical diagnosis showing that Structure-CLIP's gains on template-based attribute and relation tests transfer unevenly to host and spatial bindings in richer descriptions;
- Complex-SCLIP, a dual-positive extension that combines positive-side knowledge enhancement, shared-negative ranking, and multi-positive alignment.

2 Related Work

2.1 Compositional Evaluation of Vision–Language Models

Several benchmarks test whether CLIP-style models represent composition rather than concept co-occurrence. Broad diagnostic suites probe objects, attributes, relations, and grounded linguistic phenomena: VL-CheckList (Zhao et al., 2022) organizes tests around object, attribute, and relation

understanding, while VALSE (Parcalabescu et al., 2022) constructs controlled foils for six linguistic phenomena. ARO (Yüksekgönül et al., 2023), whose VG-Attribution and VG-Relation components are also adopted by Structure-CLIP (Huang et al., 2024), isolates attribute–object binding and relation direction. SVO-Probes (Hendricks & Nematzadeh, 2021) examines verb and argument sensitivity, whereas Winoground (Thrush et al., 2022) tests word order and role assignment using two images and two captions with nearly identical vocabularies. These benchmarks show that high retrieval accuracy does not imply reliable structural discrimination.

Later work improves the evaluation controls. CREPE (Ma et al., 2023) measures systematicity and productivity, including retrieval sets with increasing compositional complexity. SugarCrepe (Hsieh et al., 2023) constructs fluent hard negatives to reduce linguistic shortcuts, while EqBen (Wang et al., 2023) tests whether similarity changes consistently under minimal visual semantic changes. SUGAR-CREPE++ (Dumpala et al., 2024) adds a semantically equivalent but lexically different second positive caption. Hard Positive Truth (Kamath et al., 2024) shows that hard-negative fine-tuning can reduce a model's acceptance of valid hard positives. Spatial benchmarks provide complementary diagnostics: VSR (Liu et al., 2023) covers 66 natural-language spatial relations, and What'sUp (Kamath et al., 2023) controls object identity while varying relative position. Our benchmark instead places attribute, relation, host, and spatial edits in multi-entity captions and changes one visually verifiable binding per instance, testing whether gains on template-based examples remain usable when non-target facts are present.

2.2 Learning with Structural Negatives and Semantic-Preserving Positives

Structural training methods differ in how they construct negatives and retain general alignment. NegCLIP (Yüksekgönül et al., 2023) generates negative captions by swapping linguistic elements and augments each batch with visually similar nearest-neighbor images. Structure-CLIP (Huang et al., 2024) uses caption scene graphs both to generate semantic negatives and to encode object–attribute and subject–relation–object triples with a Knowledge-Enhanced Encoder (KEE). SVLC (Doveh et al., 2023b) uses language-based text manipulations to teach structured concepts such as attributes, relations, states, and actions, whereas CLoVe (Castro et al., 2024) and FSC-CLIP (Oh et al., 2024) combine compositional fine-tuning with mechanisms intended to preserve zero-shot or retrieval capability. These methods primarily optimize one correct description against structural negatives.

Positive-side approaches vary the captions or their supervision. DAC (Doveh et al., 2023a) fine-tunes with dense, image-aligned captions, LaCLIP (Fan et al., 2023) uses language rewrites, and Hard Positive Truth (Kamath et al., 2024) jointly trains with hard positives and negatives. READ-CLIP (Kwon et al., 2025) reconstructs paraphrases and aligns their sentence representations, while CLIC (Peleg et al., 2025) combines multiple positive captions, a shared hard negative, and positive-invariance training. MosaiCLIP (Singh et al., 2023) decomposes a caption scene graph into partial subgraph descriptions and aligns them with the same image from coarse to fine. Training-free methods instead modify matching at inference: ComCLIP (Jiang et al., 2024) decomposes an image into subject, object, and action regions, whereas CECE (Cascante-Bonilla et al., 2025) expands a caption with NLI-derived entailments and contradictions. In contrast, Complex-SCLIP jointly processes two full meaning-equivalent captions through a shared KEE and requires both to outrank the same structural negative.

3 Problem Formulation and Preparation

3.1 Task Formulation and Evaluation

Each instance is a tuple $D_i = (I_i, P_{1,i}, P_{2,i}, N_i, c_i)$, where I_i is an image, $P_{1,i}$ is an original correct caption, $P_{2,i}$ is a meaning-preserving paraphrase, N_i is a structural negative, and c_i denotes the edited structural family. $P_{2,i}$ may change clause order, voice, or syntactic organization, but it must preserve every depicted entity, attribute, and relation direction. N_i changes one target binding while leaving the remaining visual facts unchanged.

Let $s(I, C)$ denote the image–caption similarity. For $r \in \{1, 2\}$, the pairwise and joint decisions are

$$a_i^{(r)} = \mathbb{1}[s(I_i, P_{r,i}) > s(I_i, N_i)], \qquad a_i^{\text{Both}} = a_i^{(1)} a_i^{(2)}. \tag{1}$$

Table 1: Statistics of the constructed complex-structure benchmark.

Structural family	Target binding perturbation	Train	Validation	Test	Total
C-Attr	Attribute assignment	46	10	50	106
C-Rel	Relation direction	56	12	34	102
C-Host	Host assignment	64	14	36	114
C-LR	Left–right order	62	13	40	115
C-UD	Above–below order	59	13	40	112
C-Depth	Front–behind order	63	14	40	117
Total		**350**	**76**	**240**	**666**

For a set of M instances, we report

$$\mathrm{Acc}_{P_r} = \frac{1}{M}\sum_{i=1}^{M} a_i^{(r)}, \qquad \mathrm{Acc}_{\mathrm{Both}} = \frac{1}{M}\sum_{i=1}^{M} a_i^{\mathrm{Both}}. \tag{2}$$

The $P_1 > N$ accuracy follows the pairwise protocol used in VG-Attribution and VG-Relation (Huang et al., 2024) and is used for the initial transfer diagnosis. Both is the primary dual-positive metric, where an instance is correct only when both valid captions outrank the same negative. Ties count as errors. We also retain the continuous margins $s(I_i, P_{1,i}) - s(I_i, N_i)$ and $s(I_i, P_{2,i}) - s(I_i, N_i)$ for secondary analysis. This image-to-text setting follows the ITT protocol of SUGARCREPE++ (Dumpala et al., 2024).

Appendix Figure 3 provides a complete C-Host example with its $P_1/P_2/N$ captions.

3.2 Structural Perturbation Types

We define six structural perturbations in descriptions containing multiple entities, modifiers, or relations. **C-Attr** exchanges two visible attributes between their entities. **C-Rel** reverses the subject and object of a depicted relation. **C-Host** reassigns a visible host-bound description, such as clothing, a carried object, a local action, or a state, between two identifiable hosts. **C-LR**, **C-UD**, and **C-Depth** reverse the left–right, above–below, and front–behind order of two visible entities. Each family is defined by the edited visual binding rather than a fixed sentence template. Non-target clauses remain unchanged across P_1, P_2, and N.

3.3 Data Preparation and Quality Control

Candidate images are retrieved from Visual Genome, COCO, and V-COCO using category-specific cues from captions, scene graphs, object boxes, and action labels. These cues identify likely attribute contrasts, directed relations, separate host-bound descriptions, or sufficiently clear spatial order. They are used only for retrieval; final inclusion is based on visual inspection.

Each retained instance passes independent visual review, caption validation, and duplicate control. The review verifies that the image supports both positives and contradicts the negative, while the split procedure prevents an image from appearing in more than one partition. The resulting 666 instances are divided into 350 training, 76 validation, and 240 test records. Appendix A.1 describes candidate retrieval, the two-agent visual review, caption construction, and duplicate control in full. Table 1 summarizes the distribution.

3.4 Exploratory Transfer Diagnosis

We first evaluate pretrained CLIP and the Structure-CLIP checkpoint on all 666 instances. Neither model was trained on these records. This diagnosis uses all 666 instances because the train/test split was determined after the family-level pattern was observed. The complete family-level results are reported in Appendix Table 5. Structure-CLIP improves strongly on relation direction (+30.4 points), but its gain on C-Host is limited (+4.4) and its C-LR and C-Depth accuracy is lower than CLIP (-6.1 and -1.7). Under the stricter P_2/Both measures, the overall scores increase from 41.6%/28.7% for CLIP to 54.7%/44.6% for Structure-CLIP, while the family imbalance remains. These findings motivate dual-positive training and the extension of structural supervision to host and spatial bindings.

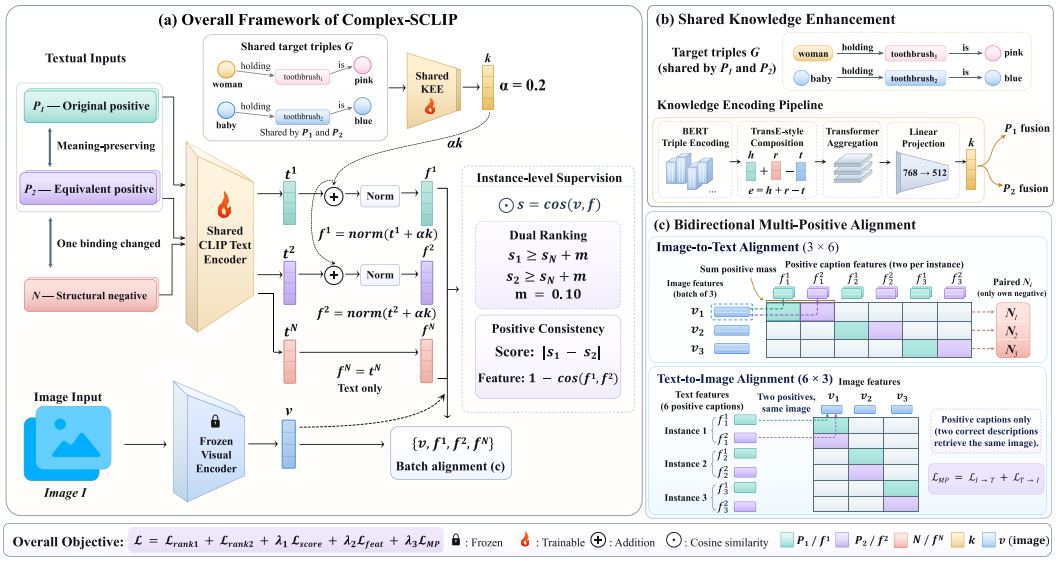

Figure 2: Complex-SCLIP framework. (a) Shared caption encoding and positive-only KEE fusion produce representations for dual ranking, positive consistency, and batch alignment with the frozen image feature. (b) The shared KEE maps the target triples of P_1 and P_2 to a common knowledge feature. (c) Bidirectional multi-positive alignment treats both captions as matches for the paired image; image-to-text additionally includes the paired structural negative, whereas text-to-image uses positive captions only.

4 Complex-SCLIP

4.1 Design Rationale and Overview

The diagnosis in Section 3 shows that a Structure-CLIP checkpoint trained with one positive caption does not transfer uniformly to host and spatial bindings. We initialize Complex-SCLIP from that checkpoint and change the learning unit from (I, P_1, N) to (I, P_1, P_2, N). The two positives express the same visual facts, while the negative changes one target binding. We train the shared text encoder to support both captions, not only the original one.

As shown in Figure 2, we encode P_1 and P_2 in the same way. Both captions share the same instance-level target triples, processed by the shared Knowledge-Enhanced Encoder (KEE), and fused into a positive representation. We rank both positives against the same negative within the instance and align them with the image across the batch. The KEE is introduced by Structure-CLIP; we apply it to two equivalent full captions and optimize the resulting representations jointly.

4.2 Dual-Positive Structural Encoding

The Knowledge-Enhanced Encoder (KEE), introduced by Structure-CLIP (Huang et al., 2024), encodes caption structure as a set of triples. Attribute pairs are written as $(\text{object}, \text{is}, \text{attribute})$, and relations retain their ordered subject, predicate, and object. Each triple (h, r, t) is tokenized and encoded by a pre-trained BERT (Devlin et al., 2019), producing embeddings $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^{768}$. We use a TransE-style (Bordes et al., 2013) composition $\mathbf{e} = \mathbf{h} + \mathbf{r} - \mathbf{t}$ to represent the relational structure, and aggregate the instance-level triples with a shared Transformer encoder. A linear projection maps the result from 768 to the CLIP feature dimension (512), yielding a structural feature k_i that is additive with the text feature.

Let t_i^1 and t_i^2 be the CLIP text features, and k_i the KEE feature computed from the shared target triples. We fuse and normalize:

$$f_i^1 = \text{norm}\big(t_i^1 + \alpha k_i\big), \qquad f_i^2 = \text{norm}\big(t_i^2 + \alpha k_i\big). \tag{3}$$

We set $\alpha = 0.2$, following the Structure-CLIP configuration. Applying the same encoder and fusion rule to both positives ensures consistent knowledge integration. The consistency losses ($\mathcal{L}_{\text{score}}$, $\mathcal{L}_{\text{feat}}$)

further encourage their outputs to agree. The negative uses the plain text feature $f_i^N = t_i^N$, matching the positive-side fusion policy of Structure-CLIP training.

4.3 Within-Instance Structural Learning

Ranking $P_1 > N$ and $P_2 > N$ independently encourages the model to resolve the structural contrast for each wording, but does not constrain the two positive representations to agree. The following losses penalize wording-dependent differences in the binding decision.

Let v_i be the frozen visual feature and $s(\cdot, \cdot)$ cosine similarity. Both positive descriptions must exceed the negative by a margin $m = 0.10$:

$$\mathcal{L}_{\text{rank1}} = \frac{1}{B} \sum_i w_i \max\big(0, m - s(v_i, f_i^1) + s(v_i, f_i^N)\big), \tag{4}$$

$$\mathcal{L}_{\text{rank2}} = \frac{1}{B} \sum_i w_i \max\big(0, m - s(v_i, f_i^2) + s(v_i, f_i^N)\big). \tag{5}$$

Here, w_i is a pre-specified category-balancing weight; its construction is given in Appendix A.2. The first loss preserves the original P_1/N learning signal. The second presents the same structural distinction through a different valid expression.

Because P_1 and P_2 denote the same visual facts, we also constrain their image-conditioned scores and features:

$$\mathcal{L}_{\text{score}} = \frac{1}{B} \sum_i w_i \left| s(v_i, f_i^1) - s(v_i, f_i^2) \right|, \tag{6}$$

$$\mathcal{L}_{\text{feat}} = \frac{1}{B} \sum_i w_i \left[1 - \cos(f_i^1, f_i^2) \right]. \tag{7}$$

These terms discourage a wording-specific decision. We evaluate the contribution of these auxiliary terms in Section 5.3.

4.4 Batch-Level Multi-Positive Alignment

The within-instance losses rank positives above the paired negative. However, they do not position them relative to other captions and images in the batch. We therefore add a batch-level contrastive term that treats both P_1 and P_2 as valid matches for the same image.

For $r \in \{1, 2\}$, define $z_{ij}^r = v_i^\top f_j^r / \tau$ and $z_{ii}^N = v_i^\top f_i^N / \tau$, with $\tau = 0.07$ and $W = \sum_i w_i$. The image-to-text term assigns positive probability mass to both paired captions and includes the paired structural negative in the denominator:

$$\mathcal{L}_{\text{I}\to\text{T}} = -\frac{1}{W} \sum_i w_i \log \frac{\exp(z_{ii}^1) + \exp(z_{ii}^2)}{\sum_j \exp(z_{ij}^1) + \sum_j \exp(z_{ij}^2) + \exp(z_{ii}^N)}. \tag{8}$$

Only N_i is used as a hard negative for image I_i, because the structural negative from another instance may be a valid description of I_i. In the text-to-image direction, each positive independently retrieves its paired image:

$$\mathcal{L}_{\text{T}\to\text{I}} = \frac{1}{2W} \sum_i w_i \left[\text{CE}_i\left((z^1)^\top, y\right) + \text{CE}_i\left((z^2)^\top, y\right) \right], \tag{9}$$

$$\mathcal{L}_{\text{MP}} = \mathcal{L}_{\text{I}\to\text{T}} + \mathcal{L}_{\text{T}\to\text{I}}. \tag{10}$$

4.5 Overall Objective and Training

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{rank1}} + \mathcal{L}_{\text{rank2}} + 0.25\mathcal{L}_{\text{score}} + 0.05\mathcal{L}_{\text{feat}} + 0.25\mathcal{L}_{\text{MP}}. \tag{11}$$

The visual encoder is frozen throughout training; the shared CLIP text encoder, the BERT encoder in KEE, the triple aggregation Transformer, and its linear projection are all updated. Optimization and checkpoint-selection details are provided in Appendix A.2.

5 EXPERIMENTS

5.1 EXPERIMENTAL PROTOCOL

We evaluate on the 240-instance test split of our controlled collection, the public SUGARCREPE++ (Dumpala et al., 2024) benchmark, and COCO 5K retrieval. Because the 240 test instances are a subset of the initial 666-instance diagnosis, we additionally evaluate on SUGARCREPE++ and COCO as external datasets.

Baselines. We compare against CLIP (Radford et al., 2021) (ViT-B/32, OpenAI weights) and Structure-CLIP (Huang et al., 2024) as the starting point; five published compositional checkpoints that share the ViT-B/32 backbone and OpenAI initialization: NegCLIP (Yüksekgönül et al., 2023), SVLC (Doveh et al., 2023b), CLoVe (Castro et al., 2024), CLIC (Peleg et al., 2025), and FSC-CLIP (Oh et al., 2024); a Structure-CLIP P1/N fine-tuning (S-CLIP P1/N FT), which continues training the same checkpoint with the released Structure-CLIP objective on our 350-instance training split; and Complex-SCLIP. This continual run measures in-domain adaptation using the released Structure-CLIP objective; Table 3 reports the controlled P_1-only comparison. All trained variants use the same data split and are optimized for 160 steps. Checkpoint availability, baseline configuration, and optimization details are given in Appendix A.2.

Metrics and significance. On the controlled test set we report P_1-accuracy, P_2-accuracy, and Both-accuracy; Both-accuracy is the primary metric. Deterministic frozen models yield a single estimate; trained models are run with three independent random seeds and reported as mean $\pm$ sample standard deviation. For the S-CLIP P1/N FT comparison, exact two-sided McNemar tests pair step-160 Complex-SCLIP with each seed's validation-selected baseline checkpoint.

The full model's optimization trajectory is shown in Appendix Figure 4.

5.2 COMPARISON WITH EXISTING METHODS

Table 2 reports P_1, P_2, and Both-accuracy on the controlled test set for all baselines.

Complex-SCLIP outperforms all baselines on Both, improving by 27.5 percentage points over the Structure-CLIP checkpoint and by 20.9 points over the in-domain S-CLIP P1/N FT baseline. Relative to S-CLIP P1/N FT, the gains are 9.1 points on P_1 and 28.6 points on P_2. The particularly large P_2 gap is consistent with explicit paraphrase supervision, while the controlled ablations below more directly assess the dual-positive contribution. The five published compositional checkpoints reach 32.1–39.6% Both, below Complex-SCLIP's 69.2% under the same protocol. Against each seed's validation-selected S-CLIP P1/N FT checkpoint, step-160 Complex-SCLIP corrects 51, 52, and 49 baseline errors; the reverse occurs in 0, 2, and 0 cases ($p < 10^{-12}$ for all exact two-sided McNemar tests). The family-level columns show the largest remaining difficulty on host and spatial bindings, although Complex-SCLIP improves both P_1 and P_2 across all six families.

5.3 COMPONENT AND OBJECTIVE ABLATIONS

We examine the contribution of each objective term through controlled ablations. All variants share the same dual-positive KEE initialisation, S-CLIP starting checkpoint, training budget, and seed set. Table 3 reports P_1, P_2, and Both-accuracy for each variant.

We first assess explicit P_2 supervision, which produces the largest change. Moving from P1-only to dual-positive ranking without $\mathcal{L}_{\text{MP}}$ raises P_1, P_2, and Both by 2.6%, 17.6%, and 13.6%, respectively. The larger increase on P_2 indicates that the second positive primarily improves robustness to a valid paraphrase. Adding $\mathcal{L}_{\text{MP}}$ contributes a further 4.8% on Both. The consistency regularizers ($\mathcal{L}_{\text{score}}$ and $\mathcal{L}_{\text{feat}}$) have limited individual effects: removing either yields 68.9–69.0% Both, within one standard deviation of the full model ($69.2 \pm 0.7\%$). We retain the pre-specified full objective as the formal model. Relative to the multi-positive model without KEE, applying the Structure-CLIP knowledge pathway to both positives raises P_1, P_2, and Both by 9.4%, 5.0%, and 8.1%.

The complete 2×2 design in Appendix Table 6 shows a KEE gain with and without $\mathcal{L}_{\text{MP}}$ and a larger $\mathcal{L}_{\text{MP}}$ gain when KEE is active. This pattern is consistent with complementary contributions from the

Table 2: Family-level and overall accuracy (%) on the 240-instance controlled test set.

(a) Original positive: $P_1 > N$

Method	Attr	Rel	Host	LR	UD	Depth	Overall
Frozen public checkpoints							
CLIP	60.0	55.9	44.4	62.5	42.5	55.0	53.8
Structure-CLIP	82.0	91.2	44.4	60.0	65.0	60.0	67.5
NegCLIP	76.0	88.2	63.9	70.0	45.0	50.0	65.4
SVLC	66.0	85.3	55.6	40.0	52.5	62.5	60.0
CLoVe	72.0	88.2	52.8	70.0	62.5	72.5	69.6
CLIC	72.0	52.9	50.0	72.5	50.0	47.5	58.3
FSC-CLIP	72.0	91.2	44.4	72.5	62.5	**75.0**	69.6
In-domain fine-tuning on 350 instances							
S-CLIP P1/N FT	77.3	88.2	47.2	72.5	62.5	65.8	69.2±1.5
Complex-SCLIP	**88.7**	**93.1**	**74.1**	**75.8**	**67.5**	70.0	**78.3±0.4**

(b) Equivalent positive: $P_2 > N$ and Both

Method	Attr	Rel	Host	LR	UD	Depth	Overall	Both
Frozen public checkpoints								
CLIP	44.0	29.4	33.3	50.0	40.0	32.5	38.8	23.8
Structure-CLIP	58.0	85.3	19.4	42.5	55.0	45.0	50.8	41.7
NegCLIP	54.0	82.4	25.0	52.5	62.5	47.5	53.8	39.6
SVLC	58.0	58.8	52.8	45.0	57.5	42.5	52.5	36.7
CLoVe	46.0	50.0	22.2	35.0	47.5	35.0	39.6	33.8
CLIC	64.0	35.3	25.0	42.5	40.0	37.5	42.1	32.1
FSC-CLIP	44.0	58.8	16.7	45.0	50.0	45.0	43.3	38.3
In-domain fine-tuning on 350 instances								
S-CLIP P1/N FT	52.7	75.5	16.7	55.8	59.2	55.0	52.5±1.1	48.3±0.4
Complex-SCLIP	**86.7**	**96.1**	**72.2**	**79.2**	**83.3**	**69.2**	**81.1±0.9**	**69.2±0.7**

Best and second-best results are shown in **bold** and underlined, respectively. Trained models report three-seed means; standard deviations are shown for aggregate metrics. The continual baseline uses P_1/N training. LR/UD/Depth denote left–right, above–below, and front–behind relations.

Table 3: Objective ablation (%) on the controlled test set. Mean ± std over three seeds. Full versus P1-only measures the complete dual-positive extension; P1/P2 without $\mathcal{L}_{\text{MP}}$ isolates the added batch objective.

Variant	P_1	P_2	Both
P1-only (no P_2 path)	76.4 ±0.2	57.1 ±0.8	50.8 ±0.7
P1/P2 without $\mathcal{L}_{\text{MP}}$	79.0 ±0.5	74.7 ±0.2	64.4 ±0.2
Multi-positive, no KEE	68.9 ±0.9	76.1 ±1.7	61.1 ±0.2
w/o $\mathcal{L}_{\text{score}}$	76.5 ±0.6	85.1 ±0.6	68.9 ±0.9
w/o $\mathcal{L}_{\text{feat}}$	77.9 ±0.4	81.7 ±1.3	69.0 ±1.3
Complex-SCLIP (full)	**78.3 ±0.4**	**81.1 ±0.9**	**69.2 ±0.7**

two components in this setting. Appendix A.4 provides a per-instance C-Host analysis of how these components affect the two ranking margins.

5.4 Public Transfer and Retrieval Preservation

We evaluate transfer on SUGARCREPE++ (Dumpala et al., 2024) and capability preservation on COCO (Lin et al., 2014) 5K retrieval. SUGARCREPE++ contains 4,757 image–caption triplets across five edit types, each pairing an image with an original positive, a valid paraphrase, and a hard negative. Unlike our multi-fact captions, it uses short descriptions and controlled phrase-level edits, thereby testing transfer to a simpler public $P_1/P_2/N$ setting. Table 4 summarizes representative checkpoints; complete results and metric definitions are reported in Appendix A.5.

Table 4: Public-benchmark summary (%). S++ denotes SUGARCREPE++; ITT/TOT are its image-conditioned/text-only triplet metrics. COCO reports Mean Recall. Trained-model values are mean $\pm$ std over three seeds.

Method	S++ P_1	S++ P_2	ITT	TOT	COCO Mean
CLIP	77.9	75.6	66.3	55.9	60.3
Structure-CLIP	83.4	74.9	69.6	69.4	64.9
CLIC	82.9	83.4	74.2	53.6	60.8
FSC-CLIP	84.6	69.8	65.9	67.6	64.8
Complex-SCLIP	$\mathbf{89.1 \pm 0.2}$	82.3 ± 1.0	$\mathbf{77.4 \pm 1.0}$	58.0 ± 1.5	61.7 ± 1.4

Complex-SCLIP achieves the highest P_1 and ITT among all compared methods. CLIC obtains the highest P_2 score (83.4%), exceeding Complex-SCLIP by 1.1%. Complex-SCLIP's TOT (58.0%) falls below the Structure-CLIP starting point (69.4%), suggesting weaker text-only discrimination; the higher ITT may reflect greater image dependence, though this alone does not confirm stronger visual grounding. Overall, dual-positive training on multi-fact descriptions transfers to shorter public edits, achieving the highest ITT among the compared methods but with reduced TOT.

COCO grouped retrieval tests whether this compositional fine-tuning preserves general image–text retrieval; no test-time KEE is applied. Complex-SCLIP's Mean Recall (61.7%) remains above CLIP (60.3%) but falls below the Structure-CLIP starting checkpoint (64.9%) and S-CLIP P1/N FT (62.3%), indicating a limited retrieval degradation from fine-tuning on the small structural dataset. Directionally, I2T R@1 improves over S-CLIP P1/N FT (43.7% versus 42.5%), while T2I R@1 drops (37.1% versus 39.2%). This asymmetry shows that the retrieval effect is not uniform across directions. FSC-CLIP, which uses an explicit retrieval-preservation objective, maintains Mean Recall at 64.8%.

6 Limitations

Scale and evaluation reuse. The 240-instance test split supports overall paired comparisons, but its 34–50 instances per family limit the precision of category-level estimates. The test records are also a subset of the diagnosis that motivated the method. SUGARCREPE++ and COCO provide independent checks, although a newly collected test set would strengthen external validity.

Text-only and retrieval trade-off. SUGARCREPE++ TOT drops from 69.4% at the Structure-CLIP initialization checkpoint to 58.0%, while COCO Mean Recall drops from 64.9% to 61.7%. These declines expose a small-data fine-tuning trade-off; isolating its cause requires controlled studies of data scale and optimization.

Model scope. All experiments use OpenAI CLIP ViT-B/32 and whole-image features. Transfer to larger backbones or other pretraining corpora remains untested, and the reported similarities do not establish region-level grounding.

7 Conclusion

In this paper, we introduce Complex-SCLIP to extend structural learning beyond single-template captions to multi-entity, multi-relation descriptions. We construct a controlled benchmark covering six structural families and find that existing structural supervision transfers unevenly, particularly to host and spatial bindings. Based on this finding, Complex-SCLIP combines equivalent positive captions, shared graph-derived knowledge, and multi-positive alignment to preserve structural decisions across valid wordings. It improves Both-accuracy by 20.9 percentage points over the in-domain Structure-CLIP P_1/N continual-fine-tuning baseline trained with the released objective and obtains the highest P_1 and ITT scores among the compared methods on SUGARCREPE++. Overall, our results support dual-positive training as an effective extension of structural learning to richer descriptions.

AI use statement

Generative AI tools were used to polish the language of selected passages and to assist with benchmark construction, including screening candidate images, assessing target bindings, and drafting selected

structural-negative captions. All AI-assisted text and annotations were reviewed and, where necessary, revised by the authors; the benchmark validation procedure is described in Appendix A.1. The authors take full responsibility for the final manuscript, dataset annotations, and reported results.

REFERENCES

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Paola Cascante-Bonilla, Yu Hou, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. Natural language inference improves compositionality in vision-language models. In *International Conference on Learning Representations*, 2025.

Santiago Castro, Amir Ziai, Avneesh Saluja, Zhuoning Yuan, and Rada Mihalcea. CLoVe: Encoding compositional language in contrastive vision-language models. *arXiv preprint arXiv:2402.15021*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Tousanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2657–2668, 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. In *Advances in Neural Information Processing Systems*, 2024.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3635–3644, 2021. doi: 10.18653/v1/2021.findings-acl.318.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2417–2425, 2024.

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Wang. ComCLIP: Training-free compositional image and text matching. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6639–6659, 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 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.568.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. The hard positive truth about vision-language compositionality. In *Computer Vision – ECCV 2024*, pp. 37–54. Springer, 2024. doi: 10.1007/978-3-031-72630-9_3.

Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Enhancing compositional reasoning in CLIP via reconstruction and alignment of text descriptions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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, 2014.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19060–19076, 2024. doi: 10.18653/v1/2024.emnlp-main.1062.

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*, pp. 8253–8280, 2022.

Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in CLIP with efficient fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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, 2021.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 869–893. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.56.

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.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11998–12008, 2023.

Mert Yüksekgönül, 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.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 30–37, 2022.

A Appendix

A.1 Benchmark Construction and Additional Diagnosis

A.1.1 Task Example

Figure 3 shows a complete C-Host instance. P_1 and P_2 express the same visual facts with different wording, whereas N swaps the clothing descriptions between the two people while preserving the remaining facts.

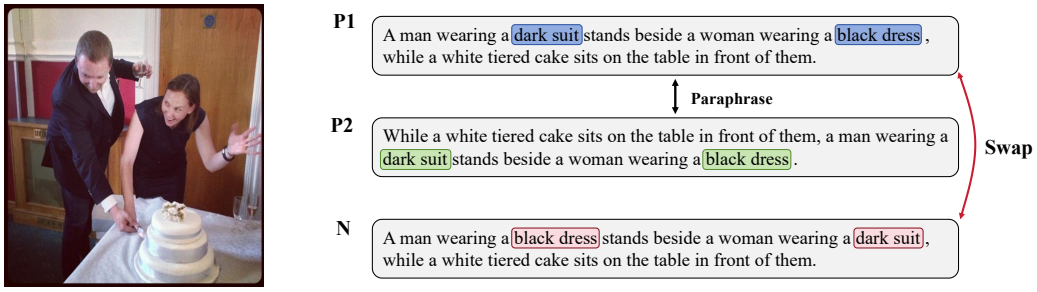

Figure 3: A C-Host instance with its original positive, equivalent positive, and structural negative captions.

A.1.2 Candidate Retrieval and Review

Candidate images are retrieved from Visual Genome, COCO, and V-COCO using category-specific cues from captions, scene graphs, object boxes, and action labels. These cues identify likely attribute contrasts, directed relations, separate host-bound descriptions, or sufficiently clear spatial order. They are used only for retrieval; eligibility is judged from the image rather than from retrieval metadata alone.

Each candidate undergoes a two-agent visual review. The primary agent identifies the visible entities and target binding and assigns a strong, borderline, or reject judgment. A second agent independently reviews the same record in a fresh session. The resulting accept, reclassify, or reject recommendations are spot-checked by an author across all six families; visually unclear or misclassified cases are revised or removed. Reclassification is allowed when another structural family provides a clearer visually verifiable edit.

Captions are written only for accepted images. P_1 describes the target binding together with non-target facts, and P_2 is a manually checked meaning-preserving rewrite. N is generated deterministically when the edit is unambiguous, as in a left–right reversal or an attribute swap. For host, state, or relation cases in which string replacement may alter the wrong phrase, a vision–language model drafts the edit and the second review checks that only the target binding changed. An instance is retained only when the image supports both positives and contradicts the negative.

A.1.3 Splits and Duplicate Control

The final benchmark contains 666 instances retained after candidate screening, visual verification, caption validation, and duplicate removal. This size follows from the quality-control process rather than a preset target. We allocate 350 instances to training, 76 to validation, and 240 to testing. Source images are grouped so that no image crosses splits. Normalized dataset names, source identifiers, and image hashes are checked jointly, while the candidate registry prevents both duplicate entries and repeated review of previously rejected images.

A.1.4 Frozen-Checkpoint Diagnosis

Table 5 reports the full family-level $P_1 > N$ diagnosis on all 666 instances. This measure is most directly comparable to the original Structure-CLIP evaluation.

Table 5: Pretrained-checkpoint $P_1 > N$ accuracy (%) on all 666 instances.

Family	n	CLIP	S-CLIP	Δ
C-Attr	106	66.0	77.4	+11.3
C-Rel	102	63.7	94.1	+30.4
C-Host	114	51.7	56.1	+4.4
C-LR	115	65.2	59.1	−6.1
C-UD	112	46.4	60.7	+14.3
C-Depth	117	63.2	61.5	−1.7
Overall	666	59.3	67.6	+8.3

Structure-CLIP improves strongly on relation direction, C-Attr, and C-UD, but its gain on C-Host is limited and its C-LR and C-Depth accuracy is lower than CLIP. Under the stricter P_2/Both measures, the overall scores increase from 41.6%/28.7% for CLIP to 54.7%/44.6% for Structure-CLIP, while the family imbalance remains.

A.2 Implementation Details and Optimization

A.2.1 Baselines and Training Configuration

CLIP (Radford et al., 2021) and Structure-CLIP (Huang et al., 2024) are the pretrained and scene-graph baselines. NegCLIP (Yüksekgönül et al., 2023) uses swapped captions and nearest-neighbor images; SVLC (Doveh et al., 2023b) uses language-structure manipulations for attributes, relations, and states; CLoVe (Castro et al., 2024) uses curated data, hard negatives, and model patching; CLIC (Peleg et al., 2025) trains with multiple images and captions; and FSC-CLIP (Oh et al., 2024) uses local hard-negative loss and selective calibrated regularization. Models without a public ViT-B/32 OpenAI-initialized checkpoint are excluded.

S-CLIP P1/N FT follows the released Structure-CLIP objective and default encoder configuration, updating both encoders with batch size 2 for at most 160 steps; the checkpoint is selected from steps 40/80/120/160 by validation Both. Complex-SCLIP uses the same fixed 160-step budget (approximately 1.8 epochs).

For the within-instance ranking losses, w_i combines the square root of inverse category frequency with a multiplier for underrepresented host and spatial families, clipped to $[0.35, 1.50]$. The visual encoder is frozen; the shared CLIP text encoder, KEE's BERT encoder, triple aggregation Transformer, and linear projection are updated. We use AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-6}, weight decay 0.1, and gradient clipping at norm 1.0.

A.2.2 Optimization Trajectory

Figure 4 shows the full model's optimization trajectory. The training objective decreases over the fixed 160-step schedule and remains low toward the end of training.

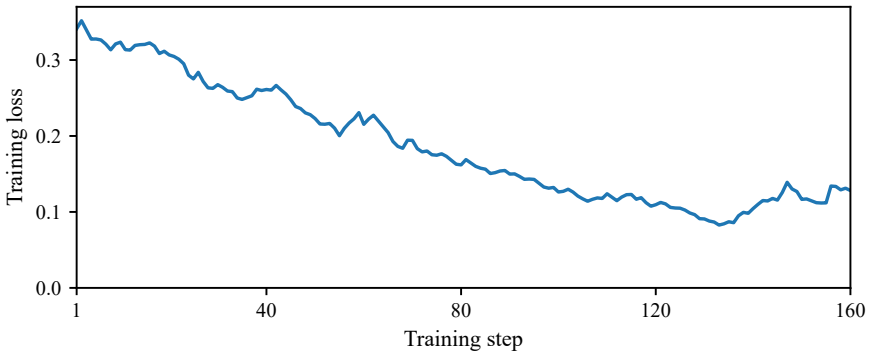

Figure 4: Optimization trajectory of Complex-SCLIP. The curve shows the exponential-moving-average smoothed mean training loss over three random seeds (smoothing factor 0.12).

A.3 Additional Factor Ablation

Table 6 presents the full 2×2 factor design crossing $\mathcal{L}_{\text{MP}}$ and KEE, with P_2 supervision always active. KEE provides a consistent main effect (+8.1% with MP and +4.0% without), while $\mathcal{L}_{\text{MP}}$ is effective mainly in the presence of KEE (+4.8% versus +0.7%). The interaction term of +4.1% indicates that the two components are complementary in this setting.

Table 6: 2×2 factor ablation (Both-accuracy, %). P_2 supervision is always active; rows vary $\mathcal{L}_{\text{MP}}$, columns vary KEE.

	+KEE	−KEE	Δ_{KEE}
$+\mathcal{L}_{\text{MP}}$	69.2 ± 0.7	61.1 ± 0.2	+8.1
$-\mathcal{L}_{\text{MP}}$	64.4 ± 0.2	60.4 ± 1.1	+4.0
Δ_{MP}	+4.8	+0.7	

A.4 Qualitative Case Analysis

Figure 5 shows a C-Host instance whose toothbrush colors must remain attached to their holders across two valid wordings. Frozen Structure-CLIP and P_1/N continual training fail P_2; adding P_2 without batch alignment leaves $\Delta_2 < 0$, and removing KEE makes Δ_1 inconsistent across seeds. Only the full model keeps both margins positive in all three runs, illustrating how dual-positive supervision, KEE, and multi-positive alignment affect the two ranking margins in this example.

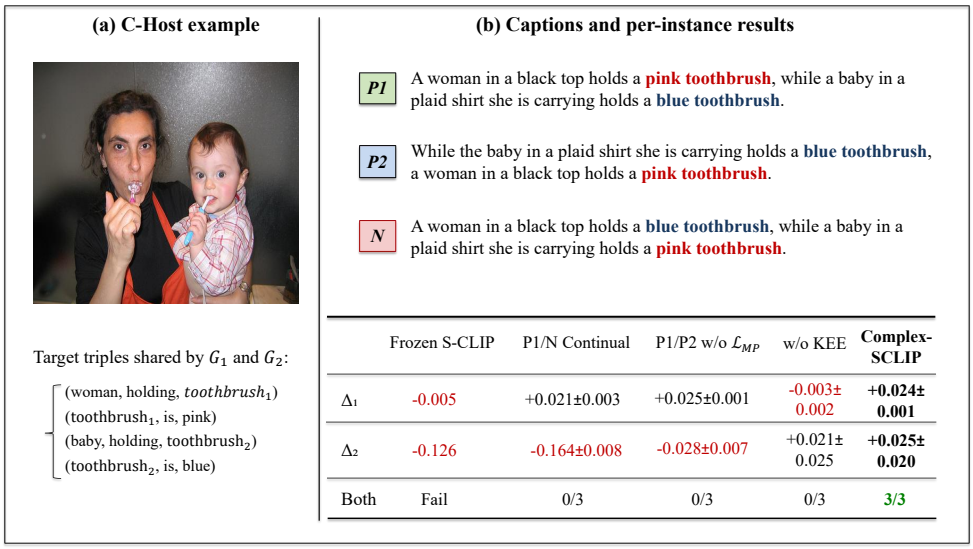

Figure 5: Per-instance comparison on a C-Host example. The negative swaps the color-to-holder bindings preserved by both positives. The table reports $\Delta_r = s(I, P_r) - s(I, N)$ over three seeds.

A.5 Full Public Transfer and Retrieval Results

This section provides the metric definitions and complete result tables for the public evaluations discussed in Section 5.4.

A.5.1 SUGARCREPE++

SUGARCREPE++ (Dumpala et al., 2024) contains 4 757 image–caption triplets across five edit types (replace, swap, add, remove, and restructure), each pairing an image with a correct caption, a lexically altered valid paraphrase, and a hard negative. ITT measures whether the model ranks the correct caption above the negative given the image, while TOT evaluates the same triplet using text

features alone. In contrast to our multi-fact captions, SUGARCREPE++ uses short descriptions and controlled phrase-level edits.

Table 7: SUGARCREPE++ (Dumpala et al., 2024) full-set results (4 757 instances, %). ITT and TOT denote image-conditioned and text-only triplet accuracy. Trained models report mean ± std over three seeds.

Method	P_1	P_2	ITT	TOT
CLIP	77.9	75.6	66.3	55.9
Structure-CLIP	83.4	74.9	69.6	69.4
NegCLIP	83.9	73.3	69.1	70.4
SVLC	75.1	71.2	60.6	63.8
CLoVe	86.1	69.5	66.0	57.9
CLIC	82.9	**83.4**	74.2	53.6
FSC-CLIP	84.6	69.8	65.9	67.6
Complex-SCLIP	$\mathbf{89.1 \pm 0.2}$	82.3 ± 1.0	$\mathbf{77.4 \pm 1.0}$	58.0 ± 1.5

A.5.2 COCO 5K RETRIEVAL

COCO (Lin et al., 2014) provides 5 000 test images with five reference captions each under the Karpathy split; all five captions for one image form a single grouped retrieval unit. I2T R@1 is the fraction of images whose ground-truth caption ranks first, T2I R@1 is the fraction of captions whose paired image ranks first, and Mean Recall averages R@1, R@5, and R@10 in both directions. No test-time KEE is applied to any trained model.

Table 8: COCO (Lin et al., 2014) 5K grouped retrieval (%). Mean Recall averages I2T and T2I R@1/5/10. Trained models report mean ± std over three seeds.

Method	I2T R@1	T2I R@1	Mean
CLIP	50.1	30.5	60.3
Structure-CLIP	48.6	39.4	64.9
CLIC	47.1	34.5	60.8
FSC-CLIP	49.7	40.4	64.8
S-CLIP P1/N FT	42.5 ± 1.5	39.2 ± 0.1	62.3 ± 0.7
Complex-SCLIP	43.7 ± 3.3	37.1 ± 0.1	$\mathbf{61.7 \pm 1.4}$