PHRASEBIND: BINDING PHRASES TO VISUAL EVIDENCE FOR FROZEN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Global image–text similarity can overlook which objects carry an attribute or participate in a relation. We introduce PhraseBind, a phrase-evidence branch that augments a frozen vision-language dual encoder. It decomposes captions into object, attribute–object, and relation phrases, routes each phrase to sparse image patches, and combines confidence-weighted local support with the unchanged global score. Counterfactual captions train binding discrimination, while phrase-region annotations supervise the supporting evidence maps. Under matched training and readout settings, PhraseBind improves ARO-VG accuracy over token- and phrase-local controls by 10.64 and 9.38 points, respectively, while retaining 99.46% of frozen-correct retrieval decisions. The learned evidence also transfers to RefCOCOg, reaching 48.5% test pointing accuracy without target-dataset supervision, compared with 34.3% for a frozen token–patch scorer. Experiments on CLIP, OpenCLIP, and SigLIP show consistent grounding gains with nearly unchanged retrieval. The same interface also complements reproduced hard-negative objectives, adding spatial evidence to their image–text matching pipelines.

1 INTRODUCTION

Dual-encoder vision-language models learn a shared image–text embedding space that supports efficient retrieval and zero-shot recognition (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). Their global similarity scores capture broad semantic compatibility, but can overlook fine-grained bindings. Distinguishing *a red cup beside a blue book* from a caption with swapped colors requires checking which object carries each attribute. Winoground, ARO, and SugarCrepe probe such failures using examples with similar visual or lexical content but different compositions (Thrush et al., 2022; Yuksekgonul et al., 2022; Hsieh et al., 2023). Figure 1 illustrates why high global compatibility can coexist with an incorrect local binding.

Hard-negative training improves binding discrimination by changing objects, attributes, or relations in captions (Yuksekgonul et al., 2022; Oh et al., 2024; Huang et al., 2024; Patel et al., 2024; Zhang et al., 2024). Local alignment methods provide complementary supervision: FILIP learns token-wise interactions, and SPARC groups image patches for token-level alignment during joint image–text pretraining (Yao et al., 2021; Bica et al., 2024). These approaches motivate a practical question for an already pretrained dual encoder: how can spatially grounded phrase evidence improve its decisions while retaining its useful global matching behavior?

We propose PhraseBind, a trainable phrase-evidence branch attached to a frozen dual encoder. It extracts object, attribute–object, and relation phrases, routes each phrase to a small set of patches, and predicts both local support and confidence in that support. Counterfactual captions teach the branch to distinguish competing bindings; grounding annotations connect its evidence maps to image regions. The global score remains fixed throughout training. At inference, confidence-weighted phrase support supplies a residual correction or a direct grounding map. Retrieval first selects candidates using cached global embeddings and then evaluates local evidence on the top candidates.

This design separates a phrase's visual support from the strength of its contribution to image–text matching. We examine that distinction with token-local and phrase-local controls trained under the same budget and evaluated with the same residual coefficients. Both controls learn strong grounding maps; confidence-weighted PhraseBind further improves ARO-VG accuracy by 9.38–10.64 points and retains more frozen-correct retrieval decisions. Thus, the relevant comparison is

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.



(a) A **statue** man sitting on a chair. (b) A **real** man sitting on a chair.

(c) A **red** cup beside a **blue** book. (d) A **blue** cup beside a **red** book.

(e) A scene with **two** cats and **one** dog on the grass. (f) A scene with **one** cat and **two** dogs on the grass.

PhraseBind	(a) → (1) ✓ , (b) → (2) ✓	(c) → (3) ✓ , (d) → (4) ✓
CLIP	(a) → (1) ✓ , (b) → (1) ✗	(c) → (3) ✓ , (d) → (3) ✗
PhraseBind	(e) → (5) ✓ , (6) → (f) ✓	
CLIP	(e) → (5) ✓ , (f) → (5) ✗	

Figure 1: Examples of binding ambiguities in global image–text matching. Phrase-level evidence can distinguish captions that share objects and scene context but assign attributes or relations differently.

between local evidence mechanisms under a shared scoring protocol. The global scorer is unchanged by construction, while preservation of the final retrieval decisions is measured empirically.

We also evaluate whether the learned evidence extends beyond its grounding supervision. An unchanged checkpoint improves RefCOCOg Test-U pointing from 34.3% to 48.5%, without using RefCOCOg annotations for training or selection. Separate runs on CLIP, OpenCLIP, and a larger SigLIP yield consistent spatial gains with nearly unchanged retrieval. Attaching the same interface to reproduced NegCLIP, TripletCLIP, and StructureCLIP objectives also improves phrase grounding and retrieval. Together, these evaluations connect the branch's spatial evidence to its use in matching: it transfers to new referring expressions, works with different frozen representations, and complements alternative hard-negative training objectives.

Our contributions are:

- **A phrase-evidence interface for frozen dual encoders.** Typed phrases, sparse routing, and confidence-weighted support provide explicit spatial evidence alongside the original global score.
- **Complementary supervision for binding and localization.** Counterfactual margins train phrase discrimination, while grounding and confidence supervision train where to gather evidence and how to weight it.
- **Controlled evaluation of local evidence and transfer.** Matched local-alignment controls isolate the empirical contribution of the interface, and independent grounding and cross-backbone experiments assess spatial transfer and retrieval preservation.

2 Related Work

Global embeddings and hard negatives. CLIP and SigLIP learn global image–text compatibility with efficient dual-tower encoders (Radford et al., 2021; Zhai et al., 2023). Compositional benchmarks reveal failures in attribute binding, relations, and word order (Thrush et al., 2022; Yuksekgonul et al., 2022; Hsieh et al., 2023; Lewis et al., 2024). NegCLIP, Structure-CLIP, TripletCLIP, CounterCurate, and FSC-CLIP improve discrimination through structured or synthetic negatives and modified training objectives (Yuksekgonul et al., 2022; Huang et al., 2024; Patel et al., 2024; Zhang et al., 2024; Oh et al., 2024). PhraseBind uses counterfactual supervision to train a separate phrase-evidence branch, keeping the pretrained global scorer fixed. Our objective-compatibility experiments attach this branch to reproduced hard-negative losses under a common frozen-backbone budget.

Local alignment and grounding. UNITER aligns words and regions, FILIP learns token-wise late interactions, and RegionCLIP and GLIP learn region-level language–image representations (Chen et al., 2020; Yao et al., 2021; Zhong et al., 2022; Li et al., 2022b). SPARC learns sparse token-conditioned patch groupings while jointly optimizing fine-grained and global pretraining objectives (Bica et al., 2024). PhraseBind shares the use of sparse local evidence, but operates on a pretrained frozen encoder: typed phrases query a separately trained local space, and confidence-weighted support is combined with the unchanged global score. This setting permits direct comparisons between token-local, phrase-local, and confidence-weighted branches without changing the backbone or readout coefficients.

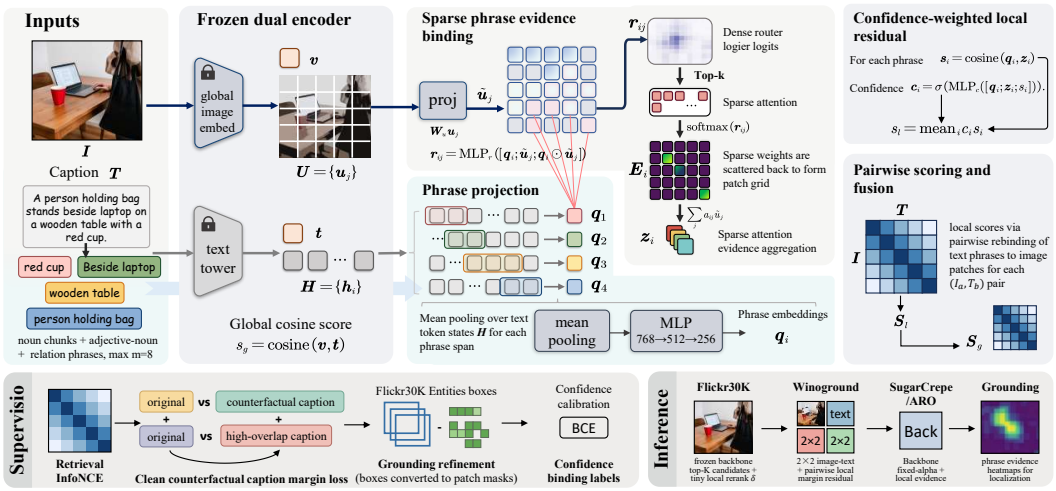

Figure 2: Framework of PhraseBind. A frozen dual encoder supplies global scores and token features. The trainable branch extracts phrases, routes them to sparse image patches, and computes confidence-weighted local support for residual scoring and grounding.

Local inference with frozen representations. Training-free localization and compositional matching methods expose information in frozen representations through local comparisons (Li et al., 2022a; Jiang et al., 2024; Miranda et al., 2026). PhraseBind learns its phrase projections, router, and confidence head using counterfactual and grounding supervision. Flickr30K Entities supplies phrase-region annotations (Plummer et al., 2015); RefCOCOg provides an independently annotated evaluation of spatial transfer (Mao et al., 2016).

3 Method

We retain the pretrained global image–text score and learn a separate branch that extracts, localizes, and weights phrase evidence (Figure 2).

3.1 Adaptive Global and Phrase-Level Evidence Fusion

The two paths provide complementary signals: the frozen encoder supplies global compatibility, and the local branch checks the visual support for individual phrases.

Frozen Global Backbone. Given an image I and a caption T, we use a pretrained dual encoder as a frozen feature extractor. It produces global embeddings $v, t \in \mathbb{R}^d$, image patch tokens $U = \{u_j\}_{j=1}^n$, and token-level text states $H = \{h_\ell\}_{\ell=1}^L$. The frozen global score is computed as the cosine similarity between normalized global embeddings:

$$s_g(I,T) = \left\langle \frac{v}{\|v\|_2}, \frac{t}{\|t\|_2} \right\rangle. \tag{1}$$

The encoder receives no gradients, so $\partial s_g / \partial \theta = 0$ for every local-branch parameter θ. All trainable components operate on the local evidence described below.

Sparse Phrase Evidence Binding. The local branch treats each textual phrase as a query for visual evidence. From caption T, we extract a set of valid phrases $\mathcal{P}(T)$, including objects, attribute–object compositions, and simple relations. For a phrase span $p_i = [a_i, b_i)$, its token states are pooled and projected into a local evidence space, $q_i = \mathrm{MLP}_\phi\left(\mathrm{pool}(\{h_\ell\}_{\ell=a_i}^{b_i-1})\right) \in \mathbb{R}^{d_\ell}$. Image patch tokens are projected to the same space as $\tilde{u}_j = W_u u_j$. This shared local space enables phrase-conditioned comparison between linguistic units and candidate visual regions.

For each phrase–patch pair, we compute a routing logit that measures how likely patch j provides evidence for phrase i:

$$r_{ij} = \mathrm{MLP}_r\left([q_i; \tilde{u}_j; q_i \odot \tilde{u}_j]\right). \tag{2}$$

Rather than attending densely to all patches, PhraseBind restricts each phrase to a sparse evidence set. Let $\mathcal{K}_i$ denote the indices of the top-k patches under r_{ij}. The selected logits are normalized to form phrase-specific attention weights, and the corresponding patch features are aggregated into an evidence vector:

$$z_i = \sum_{j \in \mathcal{K}_i} a_{ij} \tilde{u}_j, \quad \text{where} \quad a_{ij} = \frac{\exp(r_{ij})}{\sum_{j' \in \mathcal{K}_i} \exp(r_{ij'})}. \tag{3}$$

This sparse routing formulation makes the local score depend on a small set of visually accountable regions, reducing the chance that phrase evidence is diluted by background patches. The same routing logits are later supervised by grounding masks in Eq. 6, which encourages the router to represent phrase-conditioned evidence selection rather than an unconstrained cross-modal transformation.

Confidence-Weighted Local Residual. The phrase evidence score should reflect not only whether a selected region matches the phrase, but also whether the phrase admits reliable visual grounding. We therefore decompose local evidence into a support term and a reliability term. The support term s_i is the cosine similarity between the normalized phrase embedding q_i and its aggregated evidence z_i. A confidence head predicts a reliability scalar $c_i = \sigma(\mathrm{MLP}_c([q_i; z_i; s_i])) \in (0, 1)$, which is calibrated using the supervision described below. The local residual is then defined as a confidence-weighted phrase average:

$$s_\ell(I, T) = \frac{1}{|\mathcal{P}(T)|} \sum_{i \in \mathcal{P}(T)} c_i s_i. \tag{4}$$

Confidence scales the magnitude of each phrase contribution, attenuating both positive and negative support when visual grounding is uncertain.

Training Objectives. We train the phrase branch as a residual correction to the frozen global score. The fused training score is

$$s_f(I, T) = \alpha s_g(I, T) + (1 - \alpha) s_\ell(I, T). \tag{5}$$

The coefficient α controls the local contribution during training. Although s_g is fixed, the fused ranking can change; we measure retrieval preservation in Section 4.2. We apply a bidirectional InfoNCE retrieval loss $\mathcal{L}_{\mathrm{ret}}$ to the fused score matrix S_f, so that local evidence is learned under the same image–text discrimination objective used for retrieval.

To make the selected evidence spatially identifiable, we supervise the router with phrase-level grounding annotations when available. For each batch–phrase pair (b, i) with a valid mask, the normalized box-derived patch target is denoted as g_{bij}. We apply cross-entropy to the dense routing logits before top-k sparsification:

$$\mathcal{L}_{\mathrm{grd}} = -\frac{1}{|\mathcal{G}|} \sum_{(b,i) \in \mathcal{G}} \sum_{j=1}^{n} g_{bij} \log \mathrm{softmax}(r_{bi})_j, \tag{6}$$

where $\mathcal{G}$ is the subset of batch–phrase pairs containing valid masks. This supervision encourages the sparse evidence set used at inference to remain aligned with annotated visual support. In parallel, a binary confidence calibration loss $\mathcal{L}_{\mathrm{conf}}$ uses phrase-level binding labels $y_{bi}^{\mathrm{bind}} \in \{0, 1\}$ to teach the confidence head when local evidence should be trusted.

Task-Specific Inference Policies. For a given backbone, one trained checkpoint supplies the global and local scores across tasks. We use protocol-specific readouts:

$$s_{\mathrm{ret}}(I, T) = s_g(I, T) + \beta \Delta_\ell(I, T), \tag{7}$$

$$S_{\mathrm{wg}} = S_g + \beta\, \mathrm{margin}(S_\ell), \tag{8}$$

$$s_{\mathrm{comp}}(I, T) = \alpha_{\mathrm{eval}} s_g(I, T) + (1 - \alpha_{\mathrm{eval}}) s_\ell(I, T). \tag{9}$$

For large-scale retrieval, the local branch is applied as a residual re-ranking signal over globally plausible candidates. For paired compositional tests such as Winoground, the local score is converted

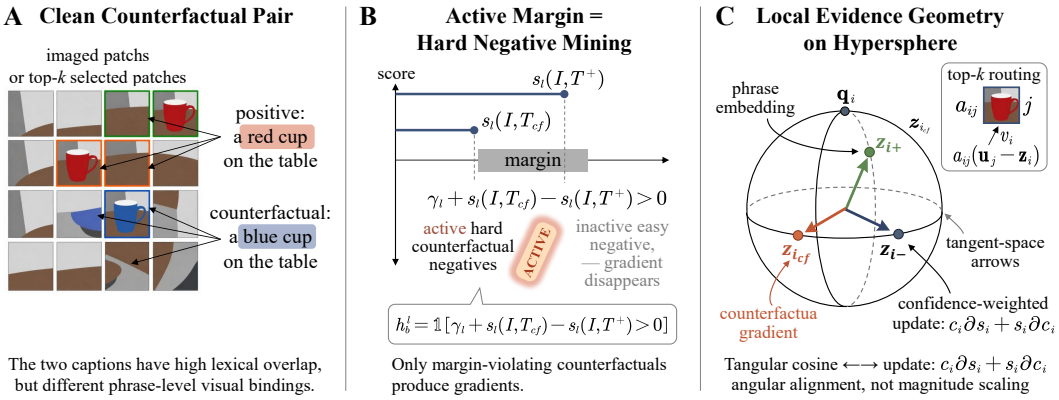

Figure 3: Clean counterfactual amplification. (A) Clean counterfactuals preserve lexical overlap but change one phrase-level visual binding. (B) The hinge margin activates gradients only for margin-violating counterfactuals. (C) Active gradients are routed through local cosine evidence and sparse top-k patch routing, strengthening phrase–patch bindings while the global backbone remains frozen.

into a pairwise margin so that evidence is compared within the ambiguous image–caption set. For binary compositional discrimination, the global and local terms are directly blended. The checkpoint is shared, while both the decision rule and residual coefficient depend on the protocol. Section 4.1 specifies coefficient selection.

3.2 Clean Counterfactual Amplification for Phrase Evidence

We further strengthen phrase evidence learning with clean counterfactual captions. Each counterfactual preserves most lexical content of the original caption but changes a specific binding, such as an attribute, relation, action, or argument. This construction makes the negative caption difficult for global semantic matching while exposing the precise local evidence that should change. As illustrated in Figure 3, the counterfactual objective activates only when the negative caption violates a margin, and its gradients are routed through the trainable phrase branch while the global backbone remains frozen.

Counterfactual Margin Objective. For each counterfactual triplet $(I_b, T_b, T_b^{\mathrm{cf}})$, we require the positive caption to score higher than its counterfactual by a margin. We impose this constraint both on the local phrase score and on the final fused score:

$$\mathcal{L}_{\mathrm{cf}}^{\ell} = \frac{1}{|\mathcal{B}_{\mathrm{cf}}|} \sum_{b \in \mathcal{B}_{\mathrm{cf}}} \max\left(0, \gamma_\ell + s_\ell(I_b, T_b^{\mathrm{cf}}) - s_\ell(I_b, T_b)\right), \tag{10}$$

$$\mathcal{L}_{\mathrm{cf}}^{f} = \frac{1}{|\mathcal{B}_{\mathrm{cf}}|} \sum_{b \in \mathcal{B}_{\mathrm{cf}}} \max\left(0, \gamma_f + s_f(I_b, T_b^{\mathrm{cf}}) - s_f(I_b, T_b)\right). \tag{11}$$

The hinge form makes the counterfactual loss selective: once a negative caption is separated from the positive caption by the required margin, it no longer contributes gradients. As a result, counterfactual training concentrates on hard negatives that remain visually plausible under the current phrase evidence branch.

The final objective combines complementary supervision signals for retrieval discrimination, counterfactual separation, spatial grounding, and confidence calibration:

$$\mathcal{L} = \lambda_{\mathrm{ret}}\mathcal{L}_{\mathrm{ret}} + \lambda_\ell\mathcal{L}_{\mathrm{cf}}^{\ell} + \lambda_f\mathcal{L}_{\mathrm{cf}}^{f} + \lambda_{\mathrm{grd}}\mathcal{L}_{\mathrm{grd}} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}}. \tag{12}$$

Learning local support and confidence. For an active counterfactual, the gradient of the local score separates into support and confidence terms:

$$\frac{\partial s_\ell}{\partial \theta} = \frac{1}{|\mathcal{P}(T)|} \sum_{i \in \mathcal{P}(T)} \left[c_i \frac{\partial s_i}{\partial \theta} + s_i \frac{\partial c_i}{\partial \theta} \right]. \tag{13}$$

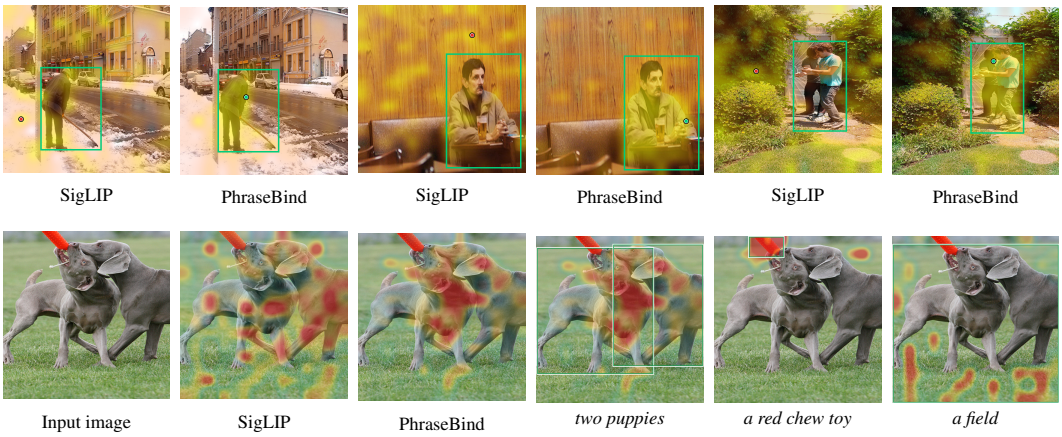

Figure 4: Visual evidence diagnostics showing PhraseBind explicitly grounds phrases to correct regions, beyond merely adjusting scalar scores. Top row: Flickr30K Entities pairs comparing frozen SigLIP and PhraseBind heatmaps. Bottom row: Dense phrase-attention decomposition for *Two puppies play with a red chew toy in a field.*, including full-caption SigLIP similarity, overall PhraseBind evidence, and individual phrase maps.

Confidence scales the support-gradient path, while the second term trains how much each phrase contributes. Counterfactual supervision updates only the local branch because $\nabla_\theta s_f = (1-\alpha)\nabla_\theta s_\ell$. Dense grounding supervision also reaches patches outside the current top-k set. Appendix A gives the cosine and routing derivatives, with the selected patch set held fixed for the sparse-attention derivative.

4 Experiments

We evaluate three aspects of PhraseBind: binding discrimination at a specified residual strength, preservation of global retrieval decisions, and transfer of the learned spatial evidence.

4.1 Experimental Setup

Training and baselines. The main model uses frozen SigLIP-B/16 at 224×224 resolution. Retrieval and counterfactual training use MSCOCO and Flickr30K captions (Lin et al., 2014; Young et al., 2014); grounding refinement uses Flickr30K Entities (Plummer et al., 2015). The development experiments use 1,000 retrieval/counterfactual updates and 300 grounding updates, followed by low-learning-rate grounding repair, with batch size 8, AdamW, and bf16. The phrase budget and local top-k are both 8. We compare against frozen SigLIP and an external CLIP ViT-B/32 reference. NegCLIP, TripletCLIP, and StructureCLIP are reproduced objectives (Yuksekgonul et al., 2022; Patel et al., 2024; Huang et al., 2024): they train global adapters under the same frozen SigLIP budget, with no phrase branch. Additional token-local and phrase-local controls share the backbone, training data, seed, schedule, and readout coefficients with PhraseBind.

Tasks and evaluation cohorts. Flickr30K retrieval uses the Karpathy test split of 1,000 images and 5,000 captions (Karpathy & Fei-Fei, 2015); we report image-to-text (I2T), text-to-image (T2I), and their mean R@1. Winoground uses the full benchmark (Thrush et al., 2022). The development comparison evaluates Flickr30K Entities pointing on 512 test samples, SugarCrepe across its available subsets (Hsieh et al., 2023), and ARO-VG attribute/relation discrimination (Yuksekgonul et al., 2022). The matched-control comparison recomputes all rows with a common split-fixed evaluator. Cross-backbone grounding uses the full canonical 5,000-caption Flickr30K Entities test split. We report these cohorts separately. All table values are percentages unless noted.

Readouts and coefficient selection. One checkpoint provides retrieval reranking, Winoground paired margins, grounding maps, and binary caption scores through Eqs. 7–9. The development coefficients were selected by sweeps using benchmark evaluation labels, without separate validation

Table 1: Development comparison using benchmark-selected residual readouts. All values are percentages; ARO-VG covers attributes and relations. F30K-Ent pointing uses 512 test samples. CLIP is an external reference. NegCLIP, TripletCLIP, and StructureCLIP are reproduced objectives using global adapters on frozen SigLIP-B/16.

Method	I2T R@1	T2I R@1	Mean R@1	WG-T	WG-I	WG-G	F30K-Ent	Sugar	ARO-VG
CLIP ViT-B/32	79.5	58.8	69.1	30.5	12.8	9.5	–	76.6	57.0
SigLIP-B/16	89.0	74.6	81.8	**33.2**	12.5	10.0	6.1	83.8	50.5
NegCLIP	86.9	73.7	80.3	31.5	10.8	8.8	6.1	**84.6**	60.7
TripletCLIP	89.1	74.0	81.5	31.8	12.0	9.8	6.1	79.6	56.6
StructureCLIP	87.6	73.4	80.5	30.5	10.8	8.8	6.1	81.8	59.6
PhraseBind	**89.4**	**74.8**	**82.1**	31.8	**15.8**	**12.0**	**63.2**	84.0	**66.6**

Table 2: Matched local-alignment controls with shared development readouts and no variant-specific coefficient tuning. All rows use the same split-fixed evaluator; these reruns are separate from Table 1. Preservation is the fraction of frozen-correct retrieval queries that remain correct. Trained variants use one seed and matched parameter budgets.

Variant	Flickr mean R@1	Preservation	F30K-Ent pointing	ARO-VG
Frozen SigLIP	81.81	100.00	6.52	50.51
Sparse token-local	81.70	99.13	63.43	56.13
Typed phrase-local	81.81	99.24	62.30	57.39
PhraseBind	81.91	99.46	65.23	66.77

splits. We label the corresponding comparisons as *development readouts*; they characterize performance at benchmark-selected operating points. Matched controls reuse these coefficients without variant-specific tuning. RefCOCOg grounding uses evidence maps directly and no target-dataset selection. Appendix C reports the post-hoc coefficient sweeps.

4.2 Main Results and Transfer

We distinguish three evaluation settings: reusing one checkpoint across datasets, training the evidence branch on other frozen backbones, and attaching it to alternative hard-negative objectives. The comparisons below report each setting with its corresponding evaluation protocol.

Overall performance. Table 1 shows nearly unchanged Flickr30K retrieval: mean R@1 increases from 81.8 to 82.1. The strongest development gains are in Flickr30K Entities pointing (6.1 to 63.2) and ARO-VG (50.5 to 66.6). Winoground image/group scores increase from 12.5/10.0 to 15.8/12.0, and SugarCrepe accuracy reaches 84.0 from 83.8. These results identify grounding and ARO-VG discrimination as the primary benefits at the selected readouts.

Matched local-alignment controls. Table 2 compares PhraseBind with token-local and typed-phrase-local branches without confidence weighting. Both controls already achieve strong pointing scores and preserve most frozen-correct retrieval decisions. PhraseBind improves pointing by 1.80 and 2.93 points and ARO-VG by 10.64 and 9.38 points over the two controls, respectively, under the same readout coefficients. The joint gains in pointing and ARO-VG support the use of confidence-weighted phrase evidence for both localization and binding discrimination.

Decision preservation. Among 4,621 retrieval queries correctly answered by the frozen model, PhraseBind retains 4,596 (99.46%), compared with 4,581 and 4,586 for the token- and phrase-local controls. It changes 25 correct answers to incorrect ones and corrects 31 previously incorrect answers. All 143 top-1 changes occur in the lowest tertile of the frozen model's top-1 versus top-2 score gap; the other two tertiles are unchanged. This is an observed property of the conservative retrieval readout. The 6,000-query pooled accuracy weights the two retrieval directions by their query counts and differs from the mean of I2T and T2I R@1.

Table 3: Frozen backbone → PhraseBind with a separately trained local branch per backbone (one seed each). F30K-Ent uses all 5,000 test captions and a matched frozen projected token–patch reference. Compositional scores use the task-specific development readouts.

Backbone	Resolution	Flickr mean R@1	WG Group	F30K-Ent pointing	SugarCrepe	ARO-VG
OpenAI CLIP ViT-B/32	224	67.1 → 67.3	7.3 → 7.8	19.5 → 59.5	76.3 → 77.0	57.2 → 54.5
OpenCLIP ViT-B/16	224	79.5 → 79.6	8.0 → 8.8	26.2 → 63.5	83.8 → 84.4	54.0 → 56.6
SigLIP-L/16	256	85.4 → 85.6	13.0 → 12.3	37.3 → 63.6	84.8 → 85.4	52.4 → 60.1

Table 4: Plug-in transfer to reproduced compositional baselines. All values are percentages; parentheses show changes in percentage points. Each pair uses the same frozen SigLIP-B/16 backbone and the scoring protocol in Section 4.1.

Method	Flickr I2T R@1	Flickr T2I R@1	Flickr mean R@1	WG Text	WG Image	WG Group	F30K-Ent Pointing	SugarCrepe
NegCLIP	86.9	73.7	80.3	31.5	10.8	8.8	6.1	84.6
Ours + NegCLIP	89.2 (+2.3)	74.7 (+1.0)	81.9 (+1.6)	33.2 (+1.7)	12.8 (+2.0)	10.2 (+1.4)	59.3 (+53.2)	83.9 (-0.7)
TripletCLIP	89.1	74.0	81.5	31.8	12.0	9.8	6.1	79.6
Ours + TripletCLIP	89.1 (+0.0)	74.7 (+0.7)	81.9 (+0.4)	33.2 (+1.4)	12.0 (+0.0)	10.0 (+0.2)	62.0 (+55.9)	83.9 (+4.3)
StructureCLIP	87.6	73.4	80.5	30.5	10.8	8.8	6.1	81.8
Ours + StructureCLIP	89.3 (+1.7)	74.6 (+1.2)	82.0 (+1.5)	31.8 (+1.3)	12.2 (+1.4)	9.5 (+0.7)	61.4 (+55.3)	83.8 (+2.0)

Cross-dataset grounding. On RefCOCOg UMD Val-U/Test-U (Mao et al., 2016), the unchanged model obtains 48.1/48.5 pointing accuracy, compared with 33.9/34.3 for a matched frozen projected token–patch scorer and 40.1/39.4 before grounding refinement. No RefCOCOg annotations are used for training, refinement, coefficient selection, or checkpoint selection. The 14.2-point Test-U gain shows that spatial evidence learned with Flickr30K Entities transfers to independently annotated referring expressions.

Cross-backbone transfer. Table 3 reports separately trained branches for OpenAI CLIP ViT-B/32, OpenCLIP ViT-B/16, and SigLIP-L/16 (Radford et al., 2021; Cherti et al., 2023; Zhai et al., 2023). Each run uses seed 42, 1,000 retrieval, 300 grounding, and 300 repair updates, with batch size 8 and phrase budget 8; feature dimensions and native resolution follow the backbone. Grounding gains are consistent across these single-run comparisons, and retrieval changes by at most 0.3 points. Pointing improves by 40.0, 37.3, and 26.3 points, respectively, extending the spatial benefits across encoder families and model sizes.

Plug-in transfer. Adding the local branch to reproduced NegCLIP, TripletCLIP, and StructureCLIP objectives raises Flickr30K Entities pointing from 6.1 to 59.3–62.0 (Table 4). Mean retrieval R@1 rises by 0.4–1.6 points, and Winoground Group by 0.2–1.4 points. This evaluates compatibility with training objectives within SigLIP-B/16; Table 3 separately evaluates backbone portability.

4.3 Ablation Study

Table 5 isolates the training components. Removing the entire local interface returns grounding and ARO-VG to the frozen baseline. Removing the counterfactual objective reduces ARO-VG from 66.6 to 62.6. Removing dense router grounding supervision reduces pointing from 63.2 to 30.0, while ARO-VG remains at 66.2. Thus, spatial supervision primarily improves localization, and counterfactual supervision contributes to binding discrimination. Removing confidence reduces Winoground Group from 12.0 to 8.8 and pointing from 63.2 to 61.0, showing its contribution to paired binding decisions and spatial evidence.

Table 5: Ablation study of PhraseBind. The upper block reports the full recipe and single-component ablations; the lower block details mechanism-targeted diagnostics. Values are percentages under the development readouts; F30K-Ent uses 512 test samples.

Variant	Flickr I2T R@1	Flickr T2I R@1	Flickr mean R@1	WG Text	WG Image	WG Group	F30K-Ent Pointing	SugarCrepe	ARO-VG
PhraseBind	89.4	74.8	82.1	31.8	15.8	12.0	63.2	84.0	66.6
w/o PhraseBind phrase/local interface	89.0	74.6	81.8	33.2	12.5	10.0	6.1	83.8	50.5
w/o clean counterfactual objective	89.5	74.6	82.1	28.5	12.8	9.8	59.7	83.8	62.6
w/o grounding recovery	89.4	74.5	82.0	29.5	14.5	9.5	60.3	84.2	66.1
w/o low-LR grounding repair	89.3	74.8	82.0	32.0	15.2	12.0	62.4	84.1	66.2
w/o confidence head	89.2	74.6	81.9	31.0	13.0	8.8	61.0	84.1	66.9
w/o dense router grounding supervision	89.4	74.8	82.1	32.0	14.0	9.8	30.0	84.3	66.2

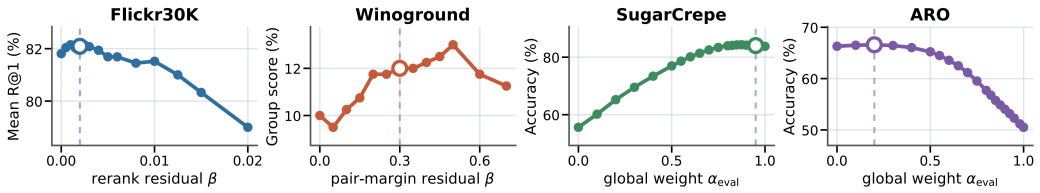

Figure 5: Post-hoc readout sensitivity on benchmark evaluation examples. All model weights are fixed; only the inference coefficient changes. Numerical values appear in Table 10.

4.4 Readout Sensitivity and Inference Cost

Readout sensitivity. Figure 5 sweeps the inference coefficient with all model weights fixed. Retrieval reaches 82.1 mean R@1 at $\beta = 0.002$ with either 10 or 20 candidates. Winoground Group increases from 11.8 to 13.0 as β rises from 0.20 to 0.50. SugarCrepe reaches 83.9–84.2 with a global-dominant blend ($\alpha_{\text{eval}} = 0.98$–$0.90$); ARO-VG reaches 65.2–66.6 with greater local weight ($\alpha_{\text{eval}} = 0.50$–$0.20$). Appendix C gives the full sweeps and coefficient-transfer evaluation.

Measured inference cost. Table 6 reports median latency over 100 queries (batch size 1, 10 warm-ups) on an NVIDIA RTX PRO 6000 Blackwell Server Edition with 1,000 GPU-cached Flickr30K images. End-to-end latency is 7.40 ms per query, including phrase extraction and local reranking, versus 3.23 ms for global retrieval. Peak allocation above the resident model and image cache is 0.113 GB. Top-K reranking restricts local scoring to $O(Kmnd_\ell)$ per query; Appendix D gives the full timing and memory protocol.

Table 6: Measured inference cost on the cached 1,000-image corpus.

Component	Median ms/query	Incremental peak GB
Frozen global retrieval	3.23	0.003
Phrase extraction	3.21	0.000
Top-10 phrase-to-patch reranking	1.60	0.112
PhraseBind end-to-end	7.40	0.113

5 Conclusion

PhraseBind adds confidence-weighted phrase-to-patch evidence to a frozen vision-language dual encoder. Counterfactual and grounding supervision train binding discrimination and spatial support. Matched controls show gains in grounding, retrieval preservation, and ARO-VG discrimination. RefCOCOg and cross-backbone evaluations demonstrate spatial transfer with nearly unchanged retrieval, while plug-in experiments establish compatibility with several hard-negative objectives.

AI USE STATEMENT

Generative AI tools were used solely to assist with and polish the writing of this manuscript. The authors reviewed and revised all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The paper and appendix specify the architecture, objectives, datasets, evaluation protocols, hyperparameters, and compute budgets. This manuscript repository includes source, figures, tables, and references. Reproducing the results also requires the public datasets and stated model versions, together with training code, evaluation scripts, checkpoints, and cached feature exports that are not included here.

REFERENCES

Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrovic. Improving fine-grained understanding in image-text pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3974–3995, 2024. URL https://proceedings.mlr.press/v235/bica24a.html.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. 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.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric 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*, 2024.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, 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 2024*, 2024.

Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022b.

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.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016. doi: 10.1109/CVPR.2016.9.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. Revisiting compositionality in dual-encoder vision-language models: The role of inference. *arXiv preprint arXiv:2604.11496*, 2026.

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.

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.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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.

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.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

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.

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.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15481–15495, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

A Detailed Derivation of Counterfactual Evidence Gradients

This section expands the gradient chain used in Section 3.2. For a single active counterfactual example, define

$$\ell_b^\ell = \gamma_\ell + s_\ell(I_b, T_b^{\mathrm{cf}}) - s_\ell(I_b, T_b), \quad h_b^\ell = \mathbf{1}[\ell_b^\ell > 0]. \tag{14}$$

where ℓ_b^ℓ is the local counterfactual hinge argument for example b; γ_ℓ is the local counterfactual margin; s_ℓ is the local phrase-evidence score; I_b is the image in example b; T_b is the original caption; T_b^{cf} is the counterfactual caption; h_b^ℓ is the binary active-margin indicator; and $\mathbf{1}[\cdot]$ is the indicator function. The local counterfactual loss can be written equivalently as the following hinge form, using the standard zero subgradient convention at the hinge boundary:

$$\mathcal{L}_{\mathrm{cf}}^\ell = \frac{1}{|\mathcal{B}_{\mathrm{cf}}|} \sum_{b \in \mathcal{B}_{\mathrm{cf}}} h_b^\ell \ell_b^\ell. \tag{15}$$

where $\mathcal{L}_{\mathrm{cf}}^\ell$ is the local counterfactual margin loss; $\mathcal{B}_{\mathrm{cf}}$ is the set of batch indices with counterfactual captions; $|\mathcal{B}_{\mathrm{cf}}|$ is the number of such indices; b indexes an example; h_b^ℓ is the active-margin indicator; and ℓ_b^ℓ is the local hinge argument. When $\ell_b^\ell \le 0$, $h_b^\ell = 0$ and the example is already margin satisfied; it contributes neither loss nor counterfactual gradient. Therefore the derivation below describes the update direction for active hard negatives, while inactive examples are handled by the same equations through the zero indicator. For any local-branch parameter θ, differentiating Eq. 15 yields

$$\nabla_\theta \mathcal{L}_{\mathrm{cf}}^\ell = \frac{1}{|\mathcal{B}_{\mathrm{cf}}|} \sum_{b \in \mathcal{B}_{\mathrm{cf}}} h_b^\ell \left[\nabla_\theta s_\ell(I_b, T_b^{\mathrm{cf}}) - \nabla_\theta s_\ell(I_b, T_b) \right]. \tag{16}$$

where $\nabla_\theta \mathcal{L}_{\mathrm{cf}}^\ell$ is the gradient of the local counterfactual loss with respect to parameter θ; θ is any trainable local-branch parameter; $\mathcal{B}_{\mathrm{cf}}$ is the counterfactual subset; b indexes an example; h_b^ℓ is the active-margin indicator; s_ℓ is the local score; I_b is the image; T_b is the original caption; and T_b^{cf} is the counterfactual caption. Thus the clean counterfactual objective does not assign a fixed weight to every negative caption. It updates the local branch only when the counterfactual caption remains within the margin of the positive caption.

For a caption with valid phrase set $\mathcal{P}$, the local score is

$$s_\ell = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} c_i s_i. \tag{17}$$

where s_ℓ is the local phrase-evidence score for the current image-caption pair; $\mathcal{P}$ is the valid phrase set for the caption; $|\mathcal{P}|$ is the number of valid phrases; i indexes a phrase; c_i is the confidence for phrase i; and s_i is the support score for phrase i. Let $\delta = \partial \mathcal{L} / \partial s_\ell$ be the scalar upstream gradient on the caption-level local score. Treating support and confidence as separate inputs to the product,

$$\frac{\partial \mathcal{L}}{\partial s_i} = \frac{\delta c_i}{|\mathcal{P}|}, \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial c_i} = \frac{\delta s_i}{|\mathcal{P}|}. \tag{19}$$

where $\partial \mathcal{L} / \partial s_i$ is the loss gradient with respect to phrase support s_i; $\partial \mathcal{L} / \partial c_i$ is the loss gradient with respect to phrase confidence c_i; $\mathcal{L}$ is the active loss being backpropagated; δ is the upstream scalar gradient on the caption score s_ℓ; and $|\mathcal{P}|$ is the number of valid phrases in the caption. When a positive caption is compared to an active counterfactual caption, δ has opposite signs for the positive and counterfactual branches: positive phrase support is pushed upward, while counterfactual phrase support is pushed downward.

The phrase support score is

$$s_i = \hat{q}_i^\top \hat{z}_i, \quad \hat{q}_i = q_i / \|q_i\|_2, \quad \hat{z}_i = z_i / \|z_i\|_2. \tag{20}$$

where s_i is the cosine support score for phrase i; q_i is the phrase embedding; z_i is the aggregated visual evidence vector; $\hat{q}_i$ is the unit-normalized phrase embedding; $\hat{z}_i$ is the unit-normalized evidence

vector; $\|\cdot\|_2$ is the Euclidean norm; and $\top$ denotes transpose. Its derivatives are

$$\frac{\partial s_i}{\partial q_i} = \frac{1}{\|q_i\|_2}(\mathbf{I} - \hat{q}_i\hat{q}_i^\top)\hat{z}_i, \tag{21}$$

$$\frac{\partial s_i}{\partial z_i} = \frac{1}{\|z_i\|_2}(\mathbf{I} - \hat{z}_i\hat{z}_i^\top)\hat{q}_i. \tag{22}$$

where $\partial s_i/\partial q_i$ is the derivative of phrase support with respect to phrase embedding q_i; $\partial s_i/\partial z_i$ is the derivative of phrase support with respect to evidence vector z_i; $\hat{q}_i = q_i/\|q_i\|_2$ and $\hat{z}_i = z_i/\|z_i\|_2$ are normalized vectors; $\mathbf{I}$ is the identity matrix; and $\top$ denotes transpose. The first derivative trains the phrase projection that produces q_i; the second trains the patch projection and router through the evidence vector z_i.

For sparse attention, $z_i = \sum_{j\in\mathcal{K}_i} a_{ij}\tilde{u}_j$ and $a_{ij} = \mathrm{softmax}_{j\in\mathcal{K}_i}(r_{ij})$. Holding the selected patch set fixed, if $g_i = \partial\mathcal{L}/\partial z_i$, the gradient with respect to a selected router logit is

$$\frac{\partial\mathcal{L}}{\partial r_{ij}} = a_{ij}\, g_i^\top(\tilde{u}_j - z_i), \quad j \in \mathcal{K}_i. \tag{23}$$

where $\partial\mathcal{L}/\partial r_{ij}$ is the loss gradient with respect to router logit r_{ij}; $\mathcal{L}$ is the active loss; a_{ij} is the sparse attention weight for phrase i and patch j; g_i is the upstream gradient with respect to evidence vector z_i; $\tilde{u}_j$ is the projected feature of patch j; z_i is the aggregated evidence vector; and $j \in \mathcal{K}_i$ means that patch j is selected in the top-k patch set for phrase i. Unselected patches have no sparse-attention gradient from the support score, but dense router logits still receive supervision from $\mathcal{L}_{\mathrm{grd}}$ in Eq. 6. This is why PhraseBind can expose sparse evidence maps at inference time while retaining a dense learning signal during grounding refinement.

The final-score counterfactual loss follows the same derivation after substituting $s_f = \alpha s_g + (1-\alpha)s_\ell$. Since the frozen backbone and identity global adapter are not trainable in the paper-facing model,

$$\nabla_\theta s_f = (1-\alpha)\nabla_\theta s_\ell \quad \text{for all local-branch parameters } \theta. \tag{24}$$

where $\nabla_\theta s_f$ is the gradient of the fused score with respect to local-branch parameter θ; s_f is the fused score; α is the fixed global residual weight; and $\nabla_\theta s_\ell$ is the gradient of the local phrase-evidence score with respect to θ. Therefore, final-score counterfactual supervision also updates the phrase evidence branch, but with the residual scale $(1-\alpha)$.

B Detailed Result Tables

Table 7 reports the provenance of each row included in the main horizontal comparison. The reproduced baseline rows are controlled reproductions under the frozen SigLIP budget, and should therefore not be interpreted as official checkpoint evaluations from the original papers.

Table 7: Result provenance for the main comparison.

Row	Status	Notes
Frozen SigLIP-B/16	frozen backbone	no trainable PhraseBind modules
NegCLIP	reproduced formal budget	global adapters only, reproduced loss
TripletCLIP	reproduced formal budget	global adapters only, reproduced loss
StructureCLIP	reproduced formal budget	global adapters only, reproduced loss
NegCLIP + PhraseBind	plug-in phrase evidence	reproduced NegCLIP objective plus local branch
TripletCLIP + PhraseBind	plug-in phrase evidence	reproduced TripletCLIP objective plus local branch
StructureCLIP + PhraseBind	plug-in phrase evidence	reproduced StructureCLIP objective plus local branch
PhraseBind-LRCF	previous balanced checkpoint	local residual, counterfactuals, task scorer
PhraseBind	final checkpoint	clean counterfactuals, recovery, low-LR grounding repair
External CLIP ViT-B/32	external reference	HuggingFace CLIP reference row

Table 8 records the development readouts for the shared checkpoint. Their coefficients were selected using benchmark evaluation labels, without separate validation splits. The protocol determines the score form; the reported scalar values are benchmark-selected operating points. Table 4 reports objective compatibility under these readouts.

Table 8: Task-specific development readouts used by PhraseBind. Scalar coefficients were selected using evaluation labels.

Task	Score readout	Key setting	Role
Flickr30K retrieval	global top-K rerank delta	$K = 10$, $\beta = 0.002$	small correction within the global candidate set
Winoground	backbone pair-delta margin	$\beta = 0.30$	expose local margins for paired hard examples
Flickr30K Entities	sparse evidence maps	no scalar blend	evaluate phrase-to-patch evidence directly
SugarCrepe	backbone/local fixed alpha	$\alpha_{\text{eval}} = 0.95$	keep global score dominant; local evidence is a small correction
ARO-VG	backbone/local fixed alpha	$\alpha_{\text{eval}} = 0.20$	emphasize local evidence for attribute/relation discrimination

Per-metric observations. The main comparison in Table 1 improves Winoground Image and Group, while Text changes from 33.2 to 31.8. SugarCrepe changes from 83.8 to 84.0, compared with 84.6 for the reproduced NegCLIP objective. In Table 3, OpenAI CLIP's ARO-VG score changes from 57.2 to 54.5, and SigLIP-L's Winoground Group score from 13.0 to 12.3. Adding the branch to NegCLIP in Table 4 changes SugarCrepe from 84.6 to 83.9. The confidence ablation in Table 5 lowers Winoground Group and pointing while increasing ARO-VG from 66.6 to 66.9. These task-specific outcomes accompany the grounding and retrieval results in the complete tables.

C Additional Experiments on Hyperparameters and Training Recipe

The paper-facing model uses one conservative local branch configuration across all tasks: phrase dimension $d_\ell = 256$, at most $m = 8$ phrases per caption, top-$k = 8$ selected patches per phrase, and fixed training fusion $\alpha = 0.98$. Table 9 records the main hyperparameters used for the development run. Readout scalars in Table 8 vary by benchmark.

Table 9: Core hyperparameters for the development PhraseBind recipe.

Component	Setting	Value	Role
Backbone	frozen encoder	SigLIP-B/16, 224×224	keep frozen unless full retraining is intended
Phrase projection	dimensions	$768 \to 512 \to 256$	use a compact local space, about one third of global dimension
Phrase budget	maximum phrases	$m = 8$	covers normal captions; increase only for long dense captions
Evidence binding	selected patches per phrase	$k = 8$	use sparse evidence, roughly one ViT neighborhood per phrase
Training fusion	global residual weight	$\alpha = 0.98$	keep training residual conservative, typically 0.95–0.99
Retrieval loss	temperature	$\tau = 0.05$	match the contrastive scale used by the frozen backbone
Clean local CF	margin, weight	$\gamma_\ell = 0.08$, weight 0.3	make local counterfactual loss visible but below retrieval scale
Clean final CF	margin, weight	$\gamma_f = 0.03$, weight 0.08	weaker penalty on the fused score
Grounding refinement	grounding, confidence weights	0.6, 0.3	prioritize dense grounding while calibrating phrase reliability
Development budget	retrieval and grounding steps	$1000 + 300$	followed by low-LR repair
Recovery repair	learning rate	10^{-5}	use low-LR repair after disabling counterfactual losses

Table 10 and Figure 5 give post-hoc sweeps on benchmark evaluation examples. Model weights are fixed and only the readout changes. These sweeps describe the sensitivity trends discussed in Section 4.4. We additionally evaluate coefficient transfer below.

Coefficient transfer. The sweeps show different preferred residual strengths: Flickr30K retrieval changes little under a small correction, while ARO-VG benefits from greater local weight. To evaluate coefficient transfer, we reuse the conservative binary setting from SugarCrepe on all 52,685 ARO-VG examples, without selecting the coefficient on ARO. Accuracy is 51.9%, compared with 66.6% at the ARO-selected operating point. This 14.7-point difference shows that checkpoint reuse alone does not determine a suitable residual coefficient for a new dataset. The benchmark-selected readouts and the fixed-coefficient transfer result therefore describe different evaluation conditions.

The leave-one-out ablation in Table 5 shows that the phrase/local interface is the main source of the grounding gain, clean counterfactual training improves compositional discrimination, and the recovery plus low-learning-rate repair stages preserve spatial evidence after stronger counterfactual supervision.

Table 10: Post-hoc sensitivity of residual readouts on benchmark evaluation data. Values are percentages. "Global" uses the frozen backbone score only. The two diagnostic columns use task-specific submetrics: I2T/T2I for Flickr30K, Text/Image for Winoground, Replace-Rel/Swap-Obj for SugarCrepe, and VG-Attr/VG-Rel for ARO-VG.

Task	Readout setting	Primary	Diagnostic 1	Diagnostic 2	Note
Flickr30K	Global	81.8	89.0	74.6	primary is mean R@1
Flickr30K	$K = 10$, $\beta = 0.002$	82.1	89.4	74.8	selected
Flickr30K	$K = 20$, $\beta = 0.002$	82.1	89.4	74.8	stable
Flickr30K	$K = 10$, $\beta = 0.010$	81.5	89.1	73.9	too strong
Winoground	$\beta = 0$	10.0	33.2	12.5	primary is WG Group
Winoground	$\beta = 0.20$	11.8	32.8	13.5	stable gain
Winoground	$\beta = 0.30$	12.0	31.8	15.8	selected
Winoground	$\beta = 0.50$	13.0	32.5	17.0	stronger residual
SugarCrepe	$\alpha_{\text{eval}} = 1.00$	83.8	70.5	60.4	primary is accuracy
SugarCrepe	$\alpha_{\text{eval}} = 0.98$	83.9	70.7	60.8	stable
SugarCrepe	$\alpha_{\text{eval}} = 0.95$	84.0	71.4	61.2	selected
SugarCrepe	$\alpha_{\text{eval}} = 0.90$	84.2	72.4	62.4	nearby best
SugarCrepe	$\alpha_{\text{eval}} = 0.70$	82.5	74.2	62.4	lower overall
ARO-VG	$\alpha_{\text{eval}} = 1.00$	50.5	54.1	46.3	primary is accuracy
ARO-VG	$\alpha_{\text{eval}} = 0.50$	65.2	72.2	56.9	local evidence helps
ARO-VG	$\alpha_{\text{eval}} = 0.30$	66.5	73.6	57.9	stable
ARO-VG	$\alpha_{\text{eval}} = 0.20$	66.6	73.6	58.2	selected
ARO-VG	$\alpha_{\text{eval}} = 0.00$	66.3	72.9	58.4	local only

D Time and Complexity Analysis

Pure dual-encoder retrieval uses cached global embeddings. For Q queries and a corpus of N candidates, applying phrase-to-patch scoring to every pair costs

$$O(QNmnd_\ell), \tag{25}$$

where m is the phrase budget, n the patch count, and d_ℓ the local dimension. Selecting the top K candidates with the global scorer reduces the local scoring cost to

$$O(QKmnd_\ell), \qquad K \ll N. \tag{26}$$

The global candidate search and query phrase extraction are additional costs. The router scores all n patches before top-k aggregation; sparsifying the evidence does not remove this dense routing computation.

Table 11: Qualitative cost comparison. The final implementation uses local evidence only where each task requires it.

Mode	Local evidence calls	Use case
Frozen global score	none after embedding cache	large-scale retrieval base
Top-K local rerank	$O(QK)$ pairs	Flickr30K retrieval
Pairwise local grid	4 pairs per example	Winoground
Single image-caption local score	positive/negative captions	SugarCrepe and ARO
Direct phrase evidence maps	matched image-caption pairs	Flickr30K Entities

Because the backbone is frozen, training memory is dominated by stored frozen features and the lightweight phrase branch rather than full backbone backpropagation. The remaining overhead comes from rebinding phrases to patches for candidate pairs and from retaining dense router logits for grounding supervision.

Measured inference overhead. We measure steady-state inference on one NVIDIA RTX PRO 6000 Blackwell Server Edition using the complete 1,000-image Flickr30K test corpus, 100 queries,

batch size 1, top-10 reranking, and 10 warm-up queries. Global image embeddings and patch tokens are cached on GPU. Table 6 reports incremental peak allocation above the 1.336 GB resident model and corpus caches. End-to-end latency is measured directly, so the component medians need not sum to its median. This protocol measures query processing with cached images and excludes initial corpus encoding.

E Reproducibility, Assets, and Responsible Use

E.1 Compute Resources

The development runs use a single-GPU frozen-backbone setting. The main trainable rows use bf16 mixed precision, batch size 8, and the shared $1000 + 300$ step budget described in Appendix C: 1,000 retrieval/counterfactual steps followed by 300 grounding-refinement steps and low-learning-rate repair. Cross-backbone runs use 300 repair updates (Section 4.2). The intended reproduction worker is one NVIDIA RTX 4090-class GPU with 24GB memory, or a compatible single GPU with at least the same memory. Frozen-backbone and external-CLIP rows require evaluation only after caching image and text features. Table 12 summarizes the compute accounting used for the reported experiments. We did not record stable wall-clock time or total GPU-hour estimates across all reported and preliminary runs; these missing time estimates are disclosed in the checklist rather than inferred from machine-dependent throughput.

Table 12: Compute resources for reproducing the reported results. The table records the reproducible hardware and step budgets, while exact wall-clock times and total GPU-hours were not logged consistently.

Experiment family	Worker and memory	Formal budget	Notes
PhraseBind final and recipe ablations	1 RTX 4090-class 24GB GPU, bf16	$1000 + 300$ steps, then repair; batch size 8	Frozen SigLIP-B/16; train only phrase projection, router, confidence, and residual modules
Plug-in NegCLIP, TripletCLIP, StructureCLIP	1 RTX 4090-class 24GB GPU, bf16	$1000 + 300$ steps, then repair; batch size 8	Same dataloaders, frozen backbone, and checkpointing stack as PhraseBind
Frozen SigLIP and external CLIP rows	1 GPU for feature extraction/evaluation	No training	Cache global features; evaluate with task-specific batch sizes
Readout sensitivity sweeps	1 GPU, inference only	No retraining	Reuse the final checkpoint and vary only scalar readout coefficients
Qualitative SDXL reranking	1 GPU for generation/scoring	Diagnostic only	Used only for figures; these samples are not benchmark metrics

E.2 Existing Assets and Licenses

The work uses existing public datasets, benchmarks, pretrained models, generated-image assets, and software packages. We cite the original papers in the main text, respect the corresponding dataset and model access terms, and do not redistribute dataset archives, gated benchmark assets, or third-party model weights as reusable assets in the anonymous paper package. Qualitative figures that derive from public benchmark examples are included only as cited low-resolution paper figures; original image rights remain with the underlying Flickr/COCO or benchmark sources. Table 13 records the principal assets, source identifiers, licenses, and terms followed in this submission.

E.3 Broader Impacts

PhraseBind is intended as a research contribution for more faithful image–text matching in frozen vision-language models. Potential positive uses include more interpretable retrieval, improved auditing of compositional VLM failures, and assistive search interfaces where phrase-level evidence

Table 13: Principal existing assets used by the paper.

Asset or identifier	Use in this paper	License or terms followed	Source
MSCOCO 2014 (Lin et al., 2014)	Retrieval training captions; COCO-backed compositional benchmark items	COCO annotations under CC BY 4.0; images remain under original Flickr licenses and terms and are not released as reusable assets	COCO
Flickr30K (Young et al., 2014) and Flickr30K Entities (Plummer et al., 2015)	Retrieval training, Flickr30K retrieval, phrase-grounding supervision/evaluation, and cited qualitative examples	Original Flickr image licenses and dataset research-use terms; paper figures are cited low-resolution examples, not a redistributed image package	F30K Entities
Winoground (Thrush et al., 2022)	Compositional evaluation	HuggingFace dataset id `facebook/winoground`; official dataset access terms; research evaluation only	HF card
SugarCrepe (Hsieh et al., 2023)	Compositional evaluation	Official repository terms; underlying COCO images follow COCO/Flickr terms	GitHub
ARO (Yuksekgonul et al., 2022)	Attribute, relation, and order evaluation	Official repository terms; underlying Visual Genome, COCO, and Flickr assets follow their original terms	GitHub
SigLIP-B/16 (Zhai et al., 2023)	Frozen backbone for all main trained rows	HuggingFace model id `google/siglip-base-patch16-224`; Apache-2.0 model-card license	HF card
CLIP ViT-B/32 (Radford et al., 2021)	External reference baseline	HuggingFace model id `openai/clip-vit-base-patch32`; MIT license for the OpenAI CLIP model/card assets used through HuggingFace	HF card
SDXL base 1.0	Qualitative candidate-pool generation only	HuggingFace model id linked at source; CreativeML Open RAIL++-M license; generated examples are diagnostic figures, not benchmark metrics	HF card
PyTorch, torchvision, HuggingFace Transformers, and HuggingFace Datasets	Training, evaluation, model loading, and dataset loading software	PyTorch/torchvision BSD-style terms and HuggingFace package Apache-2.0 terms; no modified third-party software is released in this submission	PyTorch; HF docs

helps users understand why an image was retrieved. The method may also reduce some failure modes in systems that otherwise rely only on global image-caption similarity.

The same capabilities can create risks if deployed without safeguards. Improved fine-grained retrieval may be used for surveillance, identity-linked search, or large-scale filtering of sensitive visual content. Incorrect phrase evidence can also create misleading explanations, especially for people, protected attributes, rare objects, or culturally specific visual concepts. Because the method inherits the biases and privacy limitations of pretrained vision-language backbones and public image-caption datasets, it should not be used as a standalone decision system in safety-critical, legal, employment, medical, or identity-sensitive settings. Appropriate mitigations include restricting use to consented or licensed image collections, monitoring subgroup performance, retaining human review for high-impact decisions, exposing uncertainty rather than only heatmaps, and auditing both false positives and false negatives before deployment.

F Additional Qualitative Visualizations

Figures 7 and 8 show additional samples that are not reused in the main text. These visualizations are diagnostic rather than additional benchmark results: they illustrate that PhraseBind changes the local



Figure 6: Clean diffusion candidate-pool reranking examples demonstrating PhraseBind's utility outside standard retrieval benchmarks. For each matching caption, an SDXL candidate pool is generated. We compare the candidate selected by the frozen SigLIP global score (left) with the one selected by PhraseBind evidence (right). These qualitative results illustrate how learned phrase evidence guides image selection toward precise visual bindings emphasized by the prompt.

evidence maps and can be used to select candidates with better phrase-level evidence, not only to change a scalar score.



Figure 7: Additional phrase-attention series. Each row contains seven slots: the full-caption SigLIP map, the PhraseBind evidence map, and up to five phrase-level evidence maps, with trailing slots left blank when fewer phrase maps are available. The rows show the tie/glass, girl–phone, bicycle–ramp, rainbow-art, and kids/cupcakes examples, respectively.

Figure 8 shows supplementary diffusion candidate-pool selection examples. Each candidate pool is generated first; we compare the candidate selected by frozen SigLIP with the candidate selected by PhraseBind using phrase evidence.

a red car parked beside a blue bicycle, ukiyo-e woodblock print style



SigLIP PhraseBind

A striped ceramic cup behind a small orange ball on a kitchen counter.



SigLIP PhraseBind

a green backpack on top of a brown suitcase, vintage travel poster style



SigLIP PhraseBind

two cats, one black and one white, sitting on a blue sofa, oil painting style



SigLIP PhraseBind

a man in black attire shoveling snow on a city street, loose watercolor painting style



SigLIP PhraseBind

a yellow dog sitting under a red umbrella, watercolor children's book style



SigLIP PhraseBind

Figure 8: Supplementary diffusion candidate-selection examples. The italicized text above each pair is the exact prompt used to generate and rerank that candidate pool. Each group is a two-image comparison: the left image is selected by frozen SigLIP, while the right image is selected by PhraseBind using phrase-level evidence. The expanded layout shows three comparison groups per row, without reusing the examples from the main-text qualitative figure.