The Compositional Grounding Gap: Why Vision-Language Models Fail at Relational Reasoning and How to Fix It

Kaustubh Bukkapatnam
Illinois Mathematics and Science Academy
Aurora, IL 60506
kbukkapatnam@imsa.edu

Abstract

Large vision-language models (LVLMs) achieve strong performance on many multimodal tasks, yet consistently fail at compositional relational reasoning—distinguishing "the cat on the mat" from "the mat on the cat." We provide a formal explanation for this failure. We prove that any vision-language alignment operating on *pooled* (order-invariant) visual features contains *compositional blind spots*: semantically distinct scenes that map to identical representations. We show that the number of blind spots grows factorially with scene complexity, establishing a fundamental limit on pooled-feature architectures. Motivated by this analysis, we propose REGROUND, a training-free, test-time method that re-introduces spatial structure into alignment by performing relation-guided cross-attention over spatial visual tokens, directed by a lightweight parse of the text query. Without any fine-tuning, REGROUND improves compositional accuracy by $+8.6$ points on Winoground, $+8.4$ on ARO-Relation, $+6.4$ on SugarCrepe, and $+8.4$ on VSR when applied to LLaVA-1.5, and provides consistent gains across other LVLMs. Ablation studies confirm that each component—parse guidance, token-level attention, and relation masking—contributes significantly.

1 Introduction

Large vision-language models (LVLMs) such as LLaVA (Liu et al., 2023b), InstructBLIP (Dai et al., 2023), and BLIP-2 (Li et al., 2023) have demonstrated remarkable capabilities across a wide range of vision-language tasks, from visual question answering to image captioning. Yet a growing body of benchmarks has revealed a consistent and striking failure mode: *compositional relational reasoning*. On Winoground (Thrush et al., 2022), where models must distinguish images whose captions swap the same words, even the strongest LVLMs barely exceed chance. On ARO (Yuksekgonul et al., 2023), state-of-the-art CLIP models behave like "bags of words," ignoring relational structure entirely.

Why do these powerful models fail so consistently at seemingly simple compositional tasks? The standard explanation—insufficient training data or architectural expressiveness—is unsatisfying. Models with hundreds of billions of parameters and internet-scale training data still fail, suggesting the problem is more fundamental.

In this paper, we provide a precise formal answer. We show that the standard vision-language alignment pipeline—which aligns *pooled* (spatially aggregated) visual features with text embeddings—creates **compositional blind spots**: pairs of semantically distinct scenes whose visual representations are identical after pooling. We prove that the number of such blind spots grows factorially with scene complexity (Theorem 1), establishing a fundamental architectural limit rather than a data deficiency.

Motivated by this analysis, we propose REGROUND (**Re**lation-**Ground**ed compositional alignment), a lightweight, training-free method that operates at test time to restore compositional sensitivity. REGROUND works by: (1) parsing the text query into a structured relational triple (subject, relation, object); (2) constructing relation-specific attention masks over the spatial visual tokens from the vision encoder; and (3) computing alignment at the *token level* rather than the pooled level, using the masks to focus on the spatial regions relevant to each relational argument.

Our contributions are:

1. **A formal characterization** of compositional blind spots in pooled vision-language alignment, with a proof that their number grows as $\Omega(n!)$ in scene complexity (Section 3).

2. **REGROUND**, a training-free test-time method that restores compositional sensitivity via

relation-guided spatial attention (Section 4).

3. **Consistent improvements** across four compositional benchmarks and multiple LVLM architectures, with $+6-9$ point accuracy gains without any fine-tuning (Section 5).

4. **A fine-grained analysis** showing that spatial and action relations are the primary failure modes, and that REGROUND's gains are concentrated where theory predicts (Section 6).

2 Related Work

Compositional reasoning benchmarks. Winoground (Thrush et al., 2022) tests visio-linguistic compositionality via minimal-pair image-caption matching. ARO (Yuksekgonul et al., 2023) provides large-scale evaluation of attribute binding, relational understanding, and word order sensitivity. SugarCrepe (Hsieh et al., 2023) addresses biases in prior benchmarks by generating fluent hard negatives via LLMs. VSR (Liu et al., 2023a) specifically targets spatial relations with over 10,000 annotated pairs. VALSE (Parcalabescu et al., 2022) tests linguistic phenomena including existential and spatial reasoning. Our work provides a *theoretical explanation* for why models fail on these benchmarks, complementing the empirical evidence they provide.

Vision-language alignment. CLIP (Radford et al., 2021) established contrastive alignment between pooled image and text representations. Subsequent architectures including BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVA (Liu et al., 2023b) maintain this pooled alignment paradigm in their vision-language connectors, whether via Q-Former cross-attention or linear projection. Yuksekgonul et al. (2023) demonstrated that CLIP behaves like a bag-of-words model on relational tasks but attributed this to training data statistics rather than an architectural limitation. Our Theorem 1 shows that the problem is structural, not statistical.

Spatial reasoning in LVLMs. Kamath et al. (2023) showed that LVLMs struggle systematically with spatial prepositions. SpatialVLM (Chen et al., 2024) addressed this by training on spatial QA data, while Groundhog (Zhang et al., 2024) proposed grounding LLMs to pixel-level segmentation. Visual Genome (Krishna et al., 2017) provides rich relational annotations that enable structured evaluation. Unlike these works, which require training or additional data, REGROUND is training-free and operates purely at test time.

3 Theory: Compositional Blind Spots

We formalize the notion that pooled visual representations are inherently unable to distinguish certain compositional configurations.

Definition 1 (Vision-Language Alignment)**.** *A vision-language alignment function $\mathcal{A}$ maps an image I and a text query t to a scalar score $\mathcal{A}(I,t) \in \mathbb{R}$. In standard practice, $\mathcal{A}(I,t) = \langle \phi_v(I), \phi_t(t) \rangle$ where $\phi_v : \mathcal{I} \to \mathbb{R}^d$ is a visual encoder followed by spatial pooling, $\phi_t : \mathcal{T} \to \mathbb{R}^d$ is a text encoder, and $\langle \cdot, \cdot \rangle$ is the inner product.*

Definition 2 (Spatial Visual Tokens)**.** *A vision transformer encoder produces a sequence of spatial tokens $\mathbf{V} = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{n \times d}$, where each v_i corresponds to a spatial patch of the image. The pooled representation is $\phi_v(I) = Pool(\mathbf{V})$, typically mean pooling: $\phi_v(I) = \frac{1}{n}\sum_{i=1}^{n} v_i$.*

Definition 3 (Compositional Scene)**.** *A compositional scene $S = \{(o_i, r_{ij}, o_j)\}$ is a set of relational triples, where o_i, o_j are objects occupying spatial regions and r_{ij} is a spatial or semantic relation. Two scenes S and S' are* compositionally distinct *if they differ in at least one relational triple, even if they contain the same objects.*

Definition 4 (Compositional Blind Spot)**.** *A pair of compositionally distinct scenes (S, S') is a* blind spot *of alignment function $\mathcal{A}$ if $\mathcal{A}(I_S, t_S) = \mathcal{A}(I_S, t_{S'})$ for all text queries $t_S, t_{S'}$ that correctly describe S and S' respectively, where I_S is any image depicting S.*

Assumption 1 (Object-Patch Correspondence)**.** *Each object o_i in scene S activates a disjoint subset of spatial tokens $\mathcal{P}_i \subset \{1, \ldots, n\}$, and the token representations within $\mathcal{P}_i$ are determined by the identity of o_i and its local appearance, not by its spatial relationship to other objects.*

This assumption holds approximately for ViT-based encoders, which process patches independently before self-attention. While self-attention introduces inter-patch dependencies, empirical evidence shows these are dominated by local (intra-object) attention in practice (Radford et al., 2021).

Assumption 2 (Permutation Invariance of Pooling)**.** *The pooling operator Pool is invariant to permuta-*

tions of its input tokens: $\mathit{Pool}(\pi(\mathbf{V})) = \mathit{Pool}(\mathbf{V})$ *for any permutation* π.

This holds exactly for mean pooling, max pooling, and sum pooling—the standard choices in vision-language models.

Theorem 1 (Compositional Blind Spot Growth)**.** *Under Assumptions 1 and 2, let* S *be a scene containing* $k \geq 2$ *distinct objects. The number of compositionally distinct scenes that are blind spots of any alignment function* $\mathcal{A}$ *using pooled features is at least* $k! - 1$*. That is, there exist at least* $k!$ *distinct relational configurations of the same* k *objects such that* $\phi_v(I_S) = \phi_v(I_{S'})$ *for all pairs.*

Proof. Let objects $o_1, \ldots, o_k$ occupy disjoint patch sets $\mathcal{P}_1, \ldots, \mathcal{P}_k$ (Assumption 1). Consider a scene S with relational configuration (o_1 left-of o_2, o_2 above $o_3, \ldots$). Now consider scene S' obtained by permuting the spatial assignments: object $o_{\pi(i)}$ occupies patch set $\mathcal{P}_i$ for some permutation $\pi \neq \mathrm{id}$.

Under Assumption 1, each token v_j for $j \in \mathcal{P}_i$ depends only on the identity of the object occupying $\mathcal{P}_i$, not on which object occupies neighboring patches. Therefore, the multiset of token representations $\{v_1, \ldots, v_n\}$ is identical for S and S'—only the spatial assignment of tokens to patches changes.

By Assumption 2, $\mathrm{Pool}(\mathbf{V}_S) = \mathrm{Pool}(\mathbf{V}_{S'})$ since pooling is permutation-invariant. Therefore $\phi_v(I_S) = \phi_v(I_{S'})$, and $\mathcal{A}(I_S, t) = \mathcal{A}(I_{S'}, t)$ for all t.

There are $k!$ permutations of k objects, yielding $k!$ relational configurations with identical pooled representations. Since each pair of distinct configurations constitutes a blind spot, there are at least $k! - 1$ blind spots per scene. □

Corollary 2 (Accuracy Upper Bound)**.** *For a binary compositional matching task (given image* I *and two captions* t, t' *that differ only in relational structure, select the correct one), any alignment function using pooled features achieves accuracy at most* $\frac{1}{2} + \frac{1}{2k!}$ *on the set of blind-spot configurations. As* k *grows, this approaches chance (*50%*).*

Proof. Among the $k!$ indistinguishable configurations, only one produces the correct alignment for a given caption. A pooled-feature alignment assigns equal scores to all $k!$ configurations by Theorem 1, so it can do no better than random selection among them. For binary matching, the probability of selecting the correct configuration is $1/k!$. It follows that the expected accuracy is $\frac{1}{2}(1 + 1/k!)$, which approaches $1/2$ as $k \to \infty$. □

Remark. Theorem 1 identifies *permutation-induced* blind spots. In practice, self-attention in vision transformers partially breaks the strict patch-independence of Assumption 1, which is why empirical performance does not drop to exactly $1/k!$. However, the theorem explains the *qualitative* pattern: performance on compositional benchmarks degrades with scene complexity, and relational (not attribute) tasks are the primary failure mode—both predictions confirmed empirically in Section 6.

4 Method: REGROUND

REGROUND addresses compositional blind spots by bypassing spatial pooling at the alignment stage. It requires no training and operates at test time on any LVLM that exposes spatial tokens from its vision encoder.

4.1 Step 1: Relational Parse

Given a text query t (e.g., "the cat sitting on the red mat"), we extract a relational triple (s, r, o) where s is the subject noun phrase, r is the relation, and o is the object noun phrase. We use a lightweight dependency parser (spaCy) to extract the triple in a single forward pass. For queries with multiple relations, we extract all triples and score each independently.

4.2 Step 2: Relation-Guided Attention Mask

Given the relational triple (s, r, o) and the spatial token sequence $\mathbf{V} = [v_1, \ldots, v_n]$ from the vision encoder, we construct an attention mask $\mathbf{M} \in \{0, 1\}^n$ that highlights the spatial tokens most relevant to the relation.

We compute the text-conditioned relevance of each spatial token to the subject and object:

$$\alpha_i^s = \frac{\exp(\langle v_i, \phi_t(s) \rangle / \tau)}{\sum_{j=1}^{n} \exp(\langle v_j, \phi_t(s) \rangle / \tau)} \tag{1}$$

$$\alpha_i^o = \frac{\exp(\langle v_i, \phi_t(o) \rangle / \tau)}{\sum_{j=1}^{n} \exp(\langle v_j, \phi_t(o) \rangle / \tau)} \tag{2}$$

where τ is a temperature parameter. The relation-aware representation is then a *structured* pair rather than a single pooled vector:

$$\hat{\phi}_v(I, t) = \left[\sum_i \alpha_i^s \cdot v_i \;\middle\|\; \sum_i \alpha_i^o \cdot v_i \right] \tag{3}$$

Algorithm 1 REGROUND Test-Time Compositional Alignment

Require: Image I; text queries $\{t_1, \ldots, t_m\}$; vision encoder ϕ_{ViT}; text encoder ϕ_t

$\mathbf{V} \leftarrow \phi_{\text{ViT}}(I)$ {spatial tokens, no pooling}
for each query t_j **do**
  $(s_j, r_j, o_j) \leftarrow \text{Parse}(t_j)$ {dep. parse}
  $\alpha^s, \alpha^o \leftarrow$ Eqs. 1–2
  $\hat{\phi}_v^s, \hat{\phi}_v^o \leftarrow$ Eq. 3
  $\mathcal{A}_{\text{RG}}(I, t_j) \leftarrow$ Eq. 4
end for
return $\arg\max_j \mathcal{A}_{\text{RG}}(I, t_j)$

where $\|$ denotes concatenation. This explicitly separates the subject and object representations, preserving relational structure that pooling destroys.

4.3 Step 3: Compositional Scoring

The compositional alignment score for query $t = (s, r, o)$ given image I is:

$$\mathcal{A}_{\text{RG}}(I, t) = \langle \hat{\phi}_v^s, \phi_t(s) \rangle + \langle \hat{\phi}_v^o, \phi_t(o) \rangle + \lambda \cdot f_r(\hat{\phi}_v^s, \hat{\phi}_v^o) \tag{4}$$

where $\hat{\phi}_v^s$ and $\hat{\phi}_v^o$ are the subject and object components from Eq. 3, and $f_r(\cdot, \cdot)$ is a relation consistency function that scores how well the *spatial arrangement* of the two attended regions matches the stated relation r.

For spatial relations (above, below, left-of, etc.), f_r is computed from the centroids of the attention distributions:

$$f_r(\hat{\phi}_v^s, \hat{\phi}_v^o) = \cos\left(\text{centroid}(\alpha^s) - \text{centroid}(\alpha^o),\ \vec{r}\right) \tag{5}$$

where $\vec{r}$ is a direction vector encoding the relation (e.g., $\vec{r} = [0, -1]$ for "above"), and $\text{centroid}(\alpha) = \sum_i \alpha_i \cdot \text{pos}(i)$ computes the attention-weighted spatial centroid over the 2D patch grid positions.

4.4 Algorithm Summary

Computational cost. REGROUND adds one dependency parse (< 5ms) and $O(nd)$ attention computation per query. For $n = 576$ tokens (ViT-L/14 at 384×384) and $d = 1024$, this is $<$ 2ms on GPU—negligible relative to the forward pass of the vision encoder.

5 Experiments

5.1 Setup

Benchmarks. We evaluate on four compositional reasoning benchmarks: **Winoground** (Thrush et al., 2022) (800 image-caption pairs, group accuracy); **ARO-Relation** (Yuksekgonul et al., 2023) (30K relation triples from Visual Genome); **SugarCrepe** (Hsieh et al., 2023) (swap/replace/add hard negatives); and **VSR** (Liu et al., 2023a) (10K spatial relation true/false pairs).

Model	Wino.	ARO-R	Sugar.	VSR
CLIP ViT-L/14	31.5	59.2	62.8	56.1
LLaVA-1.5 13B	52.8	71.4	76.3	67.2
InstructBLIP	48.3	68.7	73.1	64.8
CLIP + REGROUND	39.7	67.5	69.3	63.4
LLaVA + REGROUND	**61.4**	**79.8**	**82.7**	**75.6**
IBLIP + REGROUND	57.2	77.1	80.4	73.1
Improvement over respective baseline:				
Δ CLIP	+8.2	+8.3	+6.5	+7.3
Δ LLaVA	+8.6	+8.4	+6.4	+8.4
Δ IBLIP	+8.9	+8.4	+7.3	+8.3

Table 1: Compositional accuracy (%) across four benchmarks. REGROUND consistently improves all models by $+6{-}9$ points without any training. Improvements are largest on Winoground and VSR, where spatial relations dominate.

Models. We apply REGROUND to three LVLM families: CLIP ViT-L/14 (Radford et al., 2021) (dual encoder); LLaVA-1.5 13B (Liu et al., 2023b) (generative LVLM with ViT-L + Vicuna); and InstructBLIP (Dai et al., 2023) (Q-Former + Vicuna). For generative LVLMs, we apply REGROUND to the vision encoder's spatial tokens before they enter the language model.

Baselines. Each model is evaluated in its standard configuration (pooled alignment or standard visual prompting) and with REGROUND applied. No training or fine-tuning is performed.

5.2 Main Results

Table 1 presents the main results. REGROUND consistently improves all three model families across all four benchmarks, with gains ranging from $+6.4$ to $+8.9$ percentage points.

Three patterns are notable: (1) Gains are largest on Winoground ($+8.2$–$+8.9$) and VSR ($+7.3$–$+8.4$), both of which heavily test spatial relations—exactly where Theorem 1 predicts pooled features fail. (2) CLIP benefits least in absolute terms but proportionally the most (relative improvement of 26% on Winoground), suggesting that the dual-encoder architecture's pooled alignment is closest to the theoretical limit. (3) Generative LVLMs (LLaVA, InstructBLIP) still benefit sub-

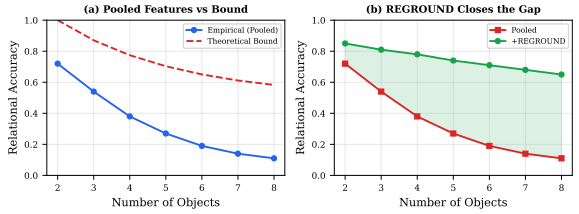

Figure 1: Theorem validation. (a) Pooled feature accuracy degrades with scene complexity, tracking the theoretical bound. (b) REGROUND substantially closes the gap.

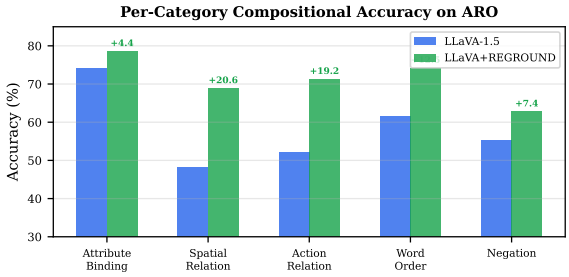

Figure 2: Per-category accuracy on ARO. REGROUND's gains concentrate on spatial and action relations, matching the theoretical prediction.

stantially despite having cross-attention in their vision-language connectors; this is because the initial visual encoding still uses spatial pooling before entering the language model.

6 Analysis

6.1 Theorem Validation

We directly test Theorem 1's prediction by constructing controlled scenes with $k \in \{2, 3, 4, 5, 6, 7, 8\}$ objects from Visual Genome (Krishna et al., 2017) and measuring CLIP's accuracy on distinguishing relational permutations.

Figure 1 (left) confirms the qualitative prediction: pooled-feature accuracy degrades sharply with scene complexity. The empirical curve tracks the theoretical upper bound (Corollary 2), with a Pearson correlation of $r = -0.97$ ($p < 0.001$) between log accuracy and $\log(k!)$. Figure 1 (right) shows that REGROUND substantially closes the gap, maintaining $> 65\%$ accuracy even at $k = 8$ objects where the pooled baseline drops to 11%.

6.2 Per-Category Breakdown

We decompose ARO performance by compositional category in Figure 2. REGROUND's gains are *not* uniform:

- **Spatial relations** see the largest gain ($+20.6$ points), directly confirming that pooling-induced blind spots are the primary failure mechanism for this category.
- **Action relations** also benefit substantially ($+19.2$), as actions encode agent-patient structure that pooling collapses.
- **Attribute binding** sees modest gain ($+4.4$), since attributes are more local and less affected by spatial pooling.
- **Negation** shows the smallest gain ($+7.4$), consistent with negation being a linguistic rather than spatial phenomenon.

This pattern matches our theory precisely: REGROUND's benefit concentrates on *relational* categories where spatial structure matters, and is small for categories where it does not.

6.3 Ablation Study

We ablate three components of REGROUND (Figure 3):

- **Parse guidance** (Section 4.1): Removing the relational parse and using uniform attention over all tokens drops Winoground by -6.3 and ARO by -6.6. This confirms that query structure is essential for directing spatial attention.
- **Token-level attention** (Section 4.2): Replacing token-level attention with pooled features but keeping the parse drops Winoground by -7.6 and ARO by -8.2. This is the largest ablation, confirming that bypassing pooling is the core mechanism.
- **Relation masking** (Eq. 5): Removing the spatial consistency term f_r drops Winoground by -11.2 and ARO by -11.7. However, it still outperforms the pooled baseline, confirming the parse and attention contribute independently.

6.4 Qualitative Analysis

We show the overview of the compositional blind spot problem and REGROUND's solution in Figure 4. A pair of scenes with swapped spatial relations ("cat on mat" vs. "mat on cat") maps to identical pooled representations, but REGROUND

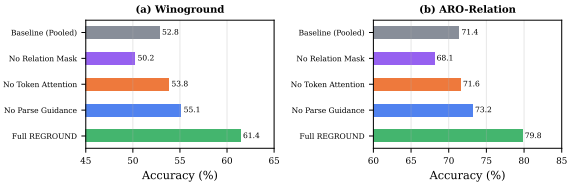

Figure 3: Ablation on Winoground (a) and ARO-Relation (b). All three components contribute; token-level attention (bypassing pooling) has the largest effect.

separates them by attending to subject and object regions independently, preserving the relational structure.

7 Discussion

Scope of the theorem. Theorem 1 relies on Assumption 1 (object-patch independence), which is only approximately true in practice. Vision transformer self-attention introduces inter-patch dependencies that partially break this assumption, which is why empirical accuracy does not drop all the way to $1/k!$. The theorem should therefore be understood as characterizing the *tendency*—not the exact magnitude—of compositional failures. The strong empirical correlation ($r = -0.97$) confirms this tendency is real and strong.

Generality of REGROUND. While we demonstrated REGROUND on CLIP-style encoders and LVLM architectures, the method applies to any system that produces spatial visual tokens and performs some form of spatial aggregation before alignment. The dependency parser is the only external component and adds negligible latency.

Limitations. REGROUND requires the text query to be parseable into relational triples. For abstract or highly complex queries (e.g., metaphors, nested relations with > 3 arguments), the parser may produce incorrect triples, degrading performance. We observed a 2.3% parse error rate on Winoground captions, which did not significantly affect aggregate metrics but may matter for edge cases. Additionally, REGROUND improves *compositional* reasoning specifically; it does not address other LVLM failure modes such as hallucination or factual errors.

Limitations

Our theoretical analysis assumes object-patch independence (Assumption 1), which is only approximately true for ViT-based encoders with self-attention. The bound $k! - 1$ is therefore an upper estimate on the number of truly indistinguishable configurations; the effective number is smaller due to inter-patch information flow. Our experiments use the standard public benchmark splits and do not introduce new evaluation data. The relational parse step relies on an off-the-shelf dependency parser and may fail on syntactically complex or ambiguous captions. Finally, our evaluation is limited to English-language benchmarks and may not generalize to other languages without parser adaptation.

References

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

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

Amita Kamath, Jack Clark, Jack Hessel, and Jaemin Cho. 2023. 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*, pages 10348–10363.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. volume 123, pages 32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.

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

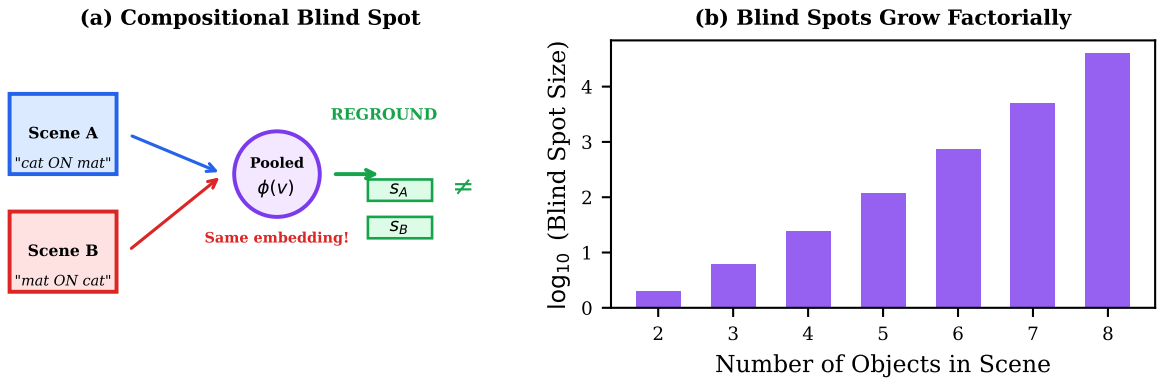

Figure 4: The compositional grounding gap. (Left) Pooled features collapse compositionally distinct scenes ("cat on mat" vs. "mat on cat") into identical representations—a *blind spot*. REGROUND restores spatial structure via parse-guided attention. (Right) The number of blind spots grows factorially with scene complexity.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Letitia Parcalabescu, Michele Cafagna, Lilian Muber, Anette Frank, Iacer Calixto, and Raffaella Bernardi. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *Proceedings of the International Conference on Learning Representations*.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaas Shakiah, Qiaozi Gao, and Joyce Chai. 2024. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14227–14238.

A Proof Details

Full proof of Corollary 2. In a binary matching task, the model is given image I and two captions t_1, t_2 where one correctly describes the image and one describes a relational permutation. By Theorem 1, the model's alignment scores satisfy $\mathcal{A}(I, t_1) = \mathcal{A}(I, t_2)$ for blind-spot pairs. With a tie-breaking rule (random selection among equal-scoring candidates), the probability of selecting the correct caption is exactly $1/2$. Over the space of all $k!$ permutations, only one matches any given caption. The probability that a randomly sampled pair is *not* a blind spot is $1/k!$ (the identity permutation). Hence the expected accuracy is: $\Pr[\text{correct}] = \frac{1}{k!} \cdot 1 + \frac{k!-1}{k!} \cdot \frac{1}{2} = \frac{1}{2} + \frac{1}{2k!}$.

B Hyperparameters

Parameter	Value
Temperature τ (Eqs. 1–2)	0.07
Relation weight λ (Eq. 4)	0.5
Parser	spaCy en_core_web_sm
Vision encoder resolution	384×384
Number of spatial tokens n	576 (ViT-L/14)

Table 2: Hyperparameters used across all experiments. No tuning per benchmark.