Diagnosing the Sources of Compositional Failure in Vision-Language Models: A Controlled Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Vision-language models (VLMs) often struggle with compositional reasoning tasks, but the reasons for this underperformance remain unclear. A common hypothesis is that models struggle to integrate multiple components, leading to training interventions to improve compositional binding. However, this assumption has never been directly quantified. Existing benchmarks evaluate captions only in their composed form, making it impossible to separate the cost of joint reasoning from the cost of recognizing individual components under increasing load. We introduce **COMPASS** (COMPositional Analysis of SkillS), a controlled evaluation framework designed to isolate and measure the distinct factors underlying compositional failure. By comparing performance on composed captions with their decomposed counterparts under matched perturbations, we directly quantify the cost of compositional integration across **87K** image-caption pairs. Across multiple VLMs, this gap is real but partial, accounting for only part of the observed degradation. This motivates a finer-grained investigation into what additional factors govern model behavior. We analyze performance at the level of individual skills: object detection, attribute binding, and relation reasoning, using skill-targeted perturbations across **274K** image-caption pairs. We find a consistent skill-specific pattern: each skill degrades primarily with the count of its own primitive type (self-load), while cross-load effects are predominantly positive, suggesting that primitives of different types provide useful grounding context. This pattern holds across standard contrastive encoders, explicitly trained compositional reasoning models, and non-contrastive architectures. These findings show that compositional degradation reflects multiple separable factors that cannot be reduced to joint reasoning alone. COMPASS provides a controlled structure for diagnosing these factors independently, enabling a more precise evaluation of how future models improve along distinct dimensions of compositional reasoning.

1 Introduction

Vision-language models (VLMs) are often deployed in settings that demand precise scene understanding, including image-text retrieval (Radford et al., 2021; Jia et al., 2021), referring expression grounding (Li et al., 2022b; Peng et al., 2023), and visual question answering (Li et al., 2022a; Liu et al., 2023), among others. In each of these settings, success requires more than recognizing individual objects. It requires understanding how objects relate to their attributes and to one another. *Compositional understanding*, the ability to understand and produce novel combinations of known concepts, is a fundamental principle of human cognition (Partee, 1984; Bottou, 2011) that neural networks have long struggled to replicate (Hupkes et al., 2020). For VLMs, compositional understanding remains a persistent and open challenge.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

Ma et al. (2023) show that model performance degrades monotonically as caption complexity increases, frequently nearing random chance at high complexity, regardless of model architecture or training dataset size. Yet despite extensive benchmarking (Yuksekgonul et al., 2023; Hsieh et al., 2023; Zhu et al., 2023), we still lack a clear answer to a basic question: what is actually failing? Most prior work has attributed compositional failure to the difficulty of integrating multiple primitives simultaneously and has focused on improving this by using training objectives, data augmentation, and architectural modifications that facilitate compositional binding (Yuksekgonul et al., 2023; Doveh et al., 2023b; Li et al., 2022b). However, existing benchmarks evaluate captions only in their composed form, without isolating the distinct factors that contribute to the observed degradation, leaving the role of joint reasoning unquantified. *To what extent does joint reasoning actually contribute to this degradation?* And if it is only a partial explanation, *what other factors govern model behavior under increasing complexity?*

To directly test these questions, we introduce **COMPASS** (COMPositional Analysis of SkillS). We construct captions from scene graphs, where each caption is composed of primitives –objects, attributes, and relations – corresponding to three core visual skills: object detection, attribute binding, and relation reasoning. By systematically varying both the type and count of these primitives, COMPASS enables two targeted analyses. First, by pairing composed captions with decomposed counterparts under matched perturbations, we isolate the **compositional integration gap**, the cost attributable to joint reasoning alone. Second, by constructing skill-targeted negatives and modeling performance as a function of per-type primitive counts, we measure **skill load**, that is, how each skill is affected by each primitive count. We further distinguish between **self-load**, degradation induced by increasing the count of the same primitive type being tested, and **cross-load**, degradation induced by increasing the count of other primitive types. In total, we evaluate compositional integration on **87K** image–caption pairs and skill load on **274K** pairs across object, attribute, and relation skills.

Across a diverse set of vision-language models spanning standard contrastive encoders, explicitly trained compositional reasoning models, and non-contrastive architectures, we observe that compositional integration incurs a consistent cost but does not fully account for the observed degradation. Performance also varies systematically with primitive counts, with each skill exhibiting sensitivity to its own components and differing behavior across primitive types. These patterns suggest that compositional failure reflects multiple interacting factors. COMPASS provides the controlled structure to diagnose these factors independently through two targeted analyses, compositional integration gap and skill load, enabling more precise evaluation of how future models improve along distinct dimensions of compositional generalization. We hope these analyses motivate skill-specific interventions that ensure individual skills remain robust as the primitive load increases.

2 Related Work

Vision-language compositionality benchmarks. Compositional understanding remains a persistent challenge for VLMs, documented across diverse phenomena: word order and caption matching (Zhu et al., 2023), spatial relations and linguistic grounding (Parcalabescu et al., 2022; Yuksekgonul et al., 2023), verb and relation understanding (Hendricks & Nematzadeh, 2021), and negation comprehension (Alhamoud et al., 2025). CREPE (Ma et al., 2023) introduces complexity-aware evaluation, establishing a monotonic degradation curve with respect to entity count that is complexity. SugarCREPE (Hsieh et al., 2023) addresses benchmark hackability by using LLM-generated negatives via single-negative retrieval. Despite this progress, existing benchmarks attribute performance degradation to joint reasoning, the cost of simultaneously composing across multiple primitive types (e.g., objects, attributes, and relations) without directly quantifying how much of the observed decline actually stems from joint reasoning. COMPASS addresses this limitation by introducing a compositional integration gap that directly measures the cost of joint reasoning under matched experimental conditions.

Several works move toward finer-grained evaluation. VL-CheckList (Zhao et al., 2022) reveals model-specific variability across primitive types, and SugarCREPE (Hsieh et al., 2023) decomposes negatives by primitive type, showing that relation negatives are harder than attribute and object negatives. However, benchmarks that vary in caption complexity treat it as a single axis, conflating the distinct contributions of individual primitive types. We address this gap by decomposing complexity into per-primitive axes to directly measure how each primitive type contributes to performance

		Evaluation Set			
	Ground Truth	**Compositional Integration**	**Skill Load**		
Structural Level	Composed	Composed/Decomposed	Object	Attribute	Relation
L3 (OAR)	47K	24K	45K	26K	34K
L2 (OA)	46K	30K	39K	37K	-
L2 (OR)	45K	33K	43K	-	34K

Table 1: **COMPASS Statistics.** We summarize the total size of our compositionality testbed across complexity levels with ground truth captions and evaluation set sizes for both compositional integration and skill load.

degradation. To understand these effects at the skill level, we construct skill-targeted negatives that perturb one primitive type at a time, isolating the load imposed independently.

Improving compositional reasoning in VLMs. A range of interventions has targeted compositional reasoning through training: hard-negative fine-tuning (Yuksekgonul et al., 2023; Zhang et al., 2024), grounding supervision (Li et al., 2022b), entity-level contrastive objectives (Doveh et al., 2023a), structured concept training (Doveh et al., 2023b), and modular encoder patching with synthetic captions (Castro et al., 2024). Training-free approaches decompose images and captions into constituent parts for local alignment (Jiang et al., 2024; Zhang et al., 2025). Despite these efforts, gains remain modest and inconsistent, and recent work shows that improvements from hard-negative fine-tuning are significantly overstated as existing benchmarks fail to probe model invariance to hard positives (Kamath et al., 2024). Hence, it is crucial to understand where improvement actually occurs, whether models become better at binding specific primitive types or at reasoning over them jointly. COMPASS provides the controlled structure needed to answer these questions, enabling precise attribution of compositional gains and failures to specific primitive types and complexity levels.

3 COMPASS: A Testbed for Analysing Compositional Failures

To diagnose the factors underlying compositional failure in VLMs as caption complexity increases, where complexity is defined as the total number of primitives (productivity in Ma et al., 2023), we introduce **COMPASS**, a controlled evaluation testbed. To enable this, we construct captions from scene graphs with explicit object, attribute, and relation structure, organizing them into hierarchical structural levels with systematic variation (Section 3.1). Models are evaluated using a retrieval-based protocol with hard negatives that isolate errors in recognizing specific primitive types (Section 3.2). This controlled structure supports two targeted analyses: **compositional integration gap** (Section 5), which isolates the cost of joint reasoning by comparing composed and decomposed captions under matched perturbations, and **skill load** (Section 6), which measures how each skill degrades based on different primitive types.

3.1 Structured Caption Construction

For controlled data construction, we use scene graphs from Visual Genome (Krishna et al., 2016), where each scene graph consists of objects (nodes), their attributes, and pairwise relations (edges) to generate captions. For each image, we sample a subgraph S using a random walk of up to ten steps (Fig. 1), starting from a random object and traversing relation edges to expand the subgraph. This ensures a coherent subset of objects and relations with sufficient primitives to construct higher-complexity captions, while maintaining a controlled set of primitives that are largely reused across captions. Including an object automatically incorporates its associated attributes. The resulting subgraph serves as a basis for generating captions across different compositional structures and complexity levels, ensuring that the same set of primitives is reused across evaluations for a given image. COMPASS is built on *5K* image – scene graph pairs from Visual Genome.

Structural Levels and Complexity Control. To disentangle how the different primitive types contribute to compositional failure, we organize captions into hierarchical structural levels. This design allows us to isolate the impact of specific skills such as object recognition, attribute binding, and relation reasoning, which are often entangled in standard benchmarks. We define captions

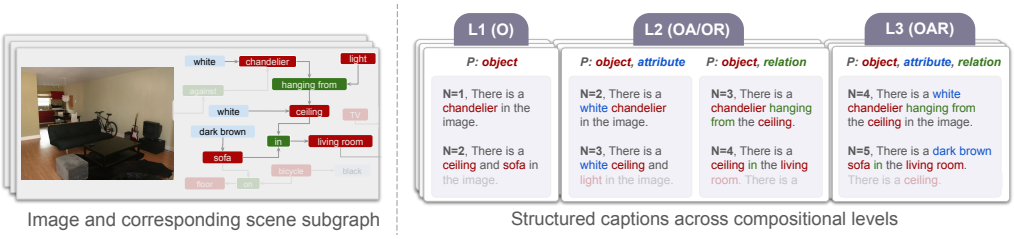


Image and corresponding scene subgraph

Structured captions across compositional levels

Figure 1: **Structured levels in COMPASS.** With a subgraph of the scene graph of an input image, we construct captions organized into structured levels based on combinations of primitives (P): L1 (O: objects), L2 (OA/OR: objects with attributes or relations), and L3 (OAR: objects, attributes, and relations). This organization enables controlled evaluation of compositional reasoning by systematically varying object (O), attribute (A), and relation (R) components across levels.

in terms of primitives corresponding to `objects (O)`, `attributes (A)`, and `relations (R)`. Structural levels are determined by the types of primitives present, as illustrated in Fig. 1: L1 includes only objects, L2 includes objects with either attributes (OA) or relations (OR), and L3 includes objects, attributes, and relations jointly (OAR). For each structural level, we vary caption complexity by controlling the total number of primitives N. Specifically, for L1, $N \in [1, 10]$; for L2 (OA), $N \in [2, 12]$; for L2 (OR), $N \in [3, 12]$; and for L3, $N \in [4, 12]$. The lower limit is defined by the minimum number of primitives of different types required to create a valid caption. As we examine the composition of different primitives, we focus on L2 and L3 in this work.

Generating Natural Language Captions. Given this structured representation, we generate natural language captions that realize these primitives across different levels and complexities. For a given complexity N and structural level defined by primitive types $t \in \{O, A, R\}$, we traverse S starting from a random node and incrementally add primitives until the caption c contains N elements, i.e., $P(c) = \{p_1, \ldots, p_N\}$. We denote the count of each primitive type by $n_t(c)$. This construction allows controlled variation in both the composition and the number of components within each caption. For the final step of synthesis, we convert the resulting primitive sets into natural language captions using GPT 4o-mini (OpenAI, 2024) with few-shot examples (see Appendix). For example, as shown in Fig. 1, "`There is a dark brown sofa in the living room. There is a ceiling.`" is an L3 caption with complexity $N = 5$, consisting of $n_o(c) = 3$ (sofa, living room, ceiling), $n_r(c) = 1$ (in), and $n_a(c) = 1$ (dark brown). This procedure yields a total of **1.38M composed ground-truth captions** across structural levels (Table 1). This structured caption space provides a controlled testbed for evaluating compositional behavior, using shared ground-truth captions across evaluations and retrieval-based evaluation using hard negatives (Section 3.2) across models.

3.2 Retrieval-based Evaluation using Hard Negatives

Compositional understanding was evaluated using an image-to-text retrieval task, where a model must identify the ground-truth caption c for an image among a set of candidate captions consisting of c and hard negatives (Hsieh et al., 2023; Ma et al., 2023; Yuksekgonul et al., 2023). Existing benchmarks have been shown to contain distributional artifacts between positive and hard-negative captions, rendering them exploitable by text-only models (Hsieh et al., 2023). Addressing this fully requires a three-stage pipeline of LLM-based generation, human validation, and adversarial refinement (Hsieh et al., 2023), which is not feasible at the scale required for COMPASS. We instead adopt an LLM-based substitution approach and conduct a perplexity audit to verify linguistic indistinguishability; full details are provided in the Appendix.

Hard Negatives Generation. Given a caption c with primitive set $P(c)$, we construct hard negatives by replacing a single primitive $p_i \in P(c)$ with another primitive of the same type t, while keeping all other primitives fixed as introduced in Shekhar et al. (2017). This results in captions that differ minimally and isolate errors in recognizing specific components. To generate replacements, we use GPT-4o mini with few-shot prompting to propose candidate primitives of the same semantic category (e.g., replacing "`sofa`" with "`chair`"). We randomly sample a replacement that does not already appear in the image's entire scene graph. We enforce additional constraints to ensure that negatives remain fluent and semantically plausible. After substitution, we adjust the caption for grammatical correctness and filter out overly similar candidates using Sentence Transformers (Reimers & Gurevych,

2019). This ensures that negatives are both natural and sufficiently distinct from the ground truth. (Refer to Appendix for more details) **Example.** For the caption "There is a dark brown sofa in the living room. There is a ceiling.", a corresponding hard negative is "There is a dark brown chair in the living room. There is a ceiling.", as shown in Figure 2.

This retrieval formulation provides the foundation for the evaluation settings introduced in Sections 5 (compositional integration) and 6 (skill load), with specific hard negatives detailed in each section.

4 Experimental Setup

To analyze compositional failure across different model families, we evaluate a diverse set of VLMs spanning standard architectures, recent variants, and models explicitly designed for compositional reasoning. Our selection includes **OpenCLIP** (Ilharco et al., 2021) (ViT-g/14, LAION-2B) as a standard contrastive baseline, along with recent variants such as **SigLIP v2** (Tschannen et al., 2025) and **PE-CLIP** (Bolya et al., 2025) (Perception Encoder). We further include compositional models: **NegCLIP** (Yuksekgonul et al., 2023), which targets compositional generalization through hard-negative training, and **CE-CLIP** (Zhang et al., 2024), which enhances compositional understanding by contrasting intra-modal and cross-modal hard negatives. We also evaluate **BLIP-L** (Li et al., 2022a), which bootstraps vision-language pretraining through a captioning and filtering pipeline and **Qwen3-VL-Embedding-8B** (Li et al., 2026), a high-capacity multimodal embedding model, to examine whether skill load effects extend beyond contrastive architectures.

All models are evaluated under a unified retrieval protocol (Section 3.2). Given an image I and a set of candidate captions, we compute image and text embeddings and rank candidates using similarity scores $s(I, c)$. This ensures a consistent evaluation setting across all architectures, enabling direct comparison of compositional integration and skill load across models. All models are evaluated on a single NVIDIA A100 GPU.

5 Compositional Integration Gap: Composed vs. Decomposed

To isolate the difficulty of joint reasoning, we compare model performance on a single composed caption against its performance on a set of independent, decomposed primitive captions. If models were perfectly compositional, performance should be invariant to whether primitives are presented jointly or in isolation. Any observed drop in the composed setting directly quantifies the cost of joint reasoning, which we define as the **compositional integration gap**.

5.1 Data Setup

Decomposed Primitive Captions. For every ground-truth caption c, we generate a set of N decomposed captions $\{d_1, \ldots, d_N\}$, where each d_i targets exactly one primitive from the original scene graph. As illustrated in Fig. 2, we define the structure of d_i based on its primitive type: objects use a simple existential caption as they can be identified by themselves (e.g., "There is a sofa in the image."); attributes use an object-attribute pair for grounding (e.g., "There is a dark brown sofa."); and relations use a triplet consisting of the relation and its two bridging objects (e.g., "The sofa is in the living room.").

Matched Perturbations for Task Equivalence. To ensure a rigorous comparison, we maintain a strict one-to-one correspondence between the negatives used in the composed and decomposed settings. For a composed caption c with complexity N, we generate N hard negatives $\{\tilde{c}_1, \ldots, \tilde{c}_N\}$, where each negative $\tilde{c}_i$ is formed by perturbing exactly one primitive $p_i \in P(c)$. Each decomposed caption d_i is then paired with a negative $\tilde{d}_i$ created using the exact same perturbation applied to its corresponding composed negative $\tilde{c}_i$. **Example.** As shown in Figure 2, if a composed caption changes the object 'sofa' to 'chair' to create a hard negative, the corresponding decomposed caption, "There is a sofa in the image." is paired with the exact same semantic perturbation "There is a chair in the image."

While a composed negative perturbs a primitive within its full context, the decomposed instance isolates that same primitive by removing the surrounding context from the prompt. To correctly retrieve the ground-truth composed caption, a model must implicitly resolve the correctness of every constituent primitive. Our decomposed setting makes this requirement explicit: success requires

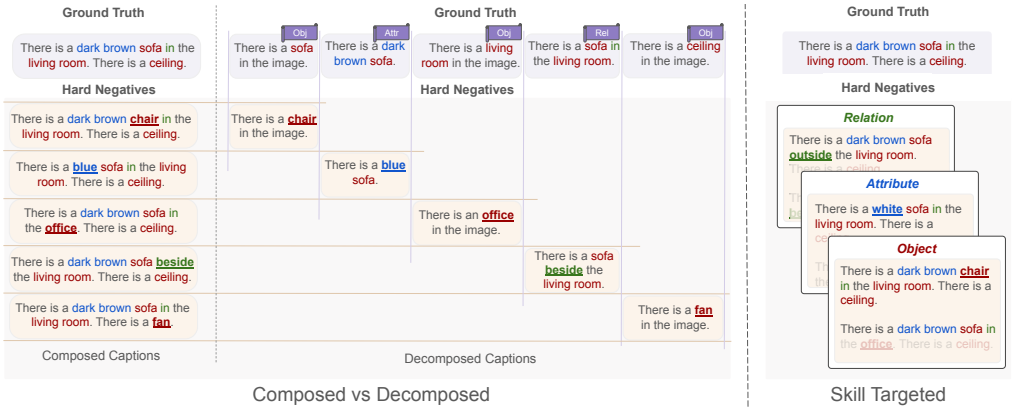


Figure 2: We construct two complementary negative structures to analyze compositional integration and skill load. **Left: Composed vs decomposed captions.** A composed caption with complexity N is paired with N hard negatives and decomposed into N primitive captions. To enable controlled comparison between joint and independent reasoning, we apply the same perturbation to create negatives for both composed and decomposed. **Right: Skill-targeted negatives.** To analyze skill load, we isolate one skill at a time and create four hard negatives by modifying that primitive type. The figure shows an example from L3 with complexity $N = 5$; this construction generalizes across all structural levels and complexities.

the model to retrieve every decomposed primitive against its respective negative independently. We construct **87K** caption pairs across structural levels to evaluate compositional integration (Table 1).

5.2 Metrics for Compositional Integration.

We formally define the **Compositional Integration Gap** (Δ) as the difference between the model's ability to retrieve the correct caption in the decomposed versus composed settings. For a composed caption c, Recall@1 is defined as:

$$\mathrm{R@1}_{\mathrm{comp}}(c) = \mathbf{1}[s(I, c) > s(I, \tilde{c}_i),\ \forall i \in \{1, \ldots, N\}]\,. \tag{1}$$

For the decomposed setting, we aggregate performance using a joint success criterion, requiring the model to correctly retrieve every single primitive in the set:

$$\mathrm{R@1}_{\mathrm{decomp}}(c) = \prod_{i=1}^{N} \mathbf{1}\left[s(I, d_i) > s(I, \tilde{d}_i)\right]. \tag{2}$$

The compositional integration gap Δ is then defined as the difference between the decomposed (Eq. 2) and composed (Eq. 1) recall scores:

$$\Delta(c) = \mathrm{R@1}_{\mathrm{decomp}}(c) - \mathrm{R@1}_{\mathrm{comp}}(c). \tag{3}$$

This formulation allows us to determine if model failure stems from the inherent complexity of joint reasoning or from the failure to recognize individual components.

5.3 How Much Does Joint Reasoning Cost?

Table 2 reports the mean and standard deviation of Δ aggregated across all complexities, with detailed results in the Appendix. Across models and structural levels, Δ is predominantly positive, confirming that joint reasoning introduces a measurable cost over independent primitive recognition.

Among standard and recent contrastive models, PE-CLIP shows the smallest integration gaps across all structural levels, suggesting greater robustness to compositional binding, while SigLIPv2 exhibits the largest gaps among this group, indicating strong individual primitive recognition that does not transfer to joint reasoning. Among non-contrastive models, BLIP-L shows moderate gaps consistent with the contrastive baseline, while Qwen3 exhibits the largest OAR gap overall, suggesting that even high-capacity multimodal embedding models struggle with compositional integration despite their representational power.

Type	OpenCLIP	SigLIPv2	PE-CLIP	NegCLIP	CE-CLIP	BLIP-L	Qwen3
OAR	0.83±0.70	2.61±1.82	1.97±0.59	3.75±0.64	-5.66±3.45	1.38±1.03	5.34±4.42
OA	2.39±1.23	12.65±2.53	0.94±1.96	-2.70±1.83	-20.27±5.17	1.19±2.17	11.94±4.81
OR	17.56±1.98	8.64±5.04	2.69±0.70	5.54±1.44	-4.96±3.93	2.30±2.93	7.31±5.16

Table 2: **Compositional Integration (Aggregated).** We report the mean ± standard deviation of the difference between decomposed and composed accuracy (Δ) aggregated across all complexities. Positive values indicate better independent than joint performance, highlighting compositional binding difficulty. Shaded cells indicate models for which composed captions are easier to retrieve than their decomposed counterparts.

NegCLIP and CE-CLIP are notable exceptions, both achieving negative Δ values, indicating that composed performance exceeds decomposed performance. NegCLIP shows this behavior only on OA, while CE-CLIP shows it consistently across all structural levels. This suggests that explicit compositional training enables these models to leverage joint context across primitives more effectively, such that composed captions are actually easier to retrieve than their decomposed counterparts. This qualitatively different behavior suggests that compositional training objectives can improve joint reasoning rather than merely shifting distributional preferences.

Discussion. The compositional integration gap provides a useful upper bound on model capability, showing how well models perform when each primitive is evaluated in isolation. While compositional training objectives such as NegCLIP and CE-CLIP can reverse the integration gap, making composed retrieval easier than decomposed retrieval, both models, like CLIP, still exhibit monotonic performance degradation with increasing caption complexity on CREPE's productivity benchmark (Figure 3, refer to the appendix for more details). This raises a deeper question: *what drives performance degradation as caption complexity grows, beyond joint reasoning alone?* Prior work has treated complexity as a single scalar, without distinguishing the contributions of individual primitive types. To answer this, we move beyond this aggregate view and analyze how model performance varies with the counts of each primitive type independently, discussed next in Section 6.

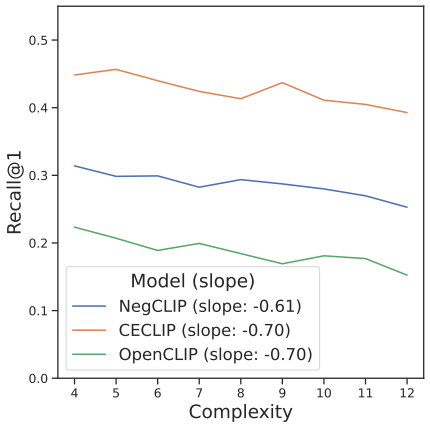


Figure 3: Performance on CREPE

6 Skill Load: Skill Targeted Perturbations

Building on the discussion above, we turn to a more fine-grained analysis of what drives performance degradation in composed captions. Specifically, we examine how performance on each skill—object, attribute, and relation—varies with the number of primitives of each type—what we term ***skill load***. We distinguish between ***self-load***, the degradation induced by increasing the count of the same primitive type being tested, and ***cross-load***, the degradation induced by increasing the count of other primitive types.

6.1 Data Setup

To isolate effects on individual skills, we construct skill-targeted hard negatives. For a given primitive type $t \in \{O, A, R\}$, we replace primitives of type t only, keeping all other primitives fixed, allowing us to measure a model's ability to detect a specific primitive type independently of perturbations to others. We generate $K = 4$ hard negatives per caption, as Udandarao et al. (2025) shows that single-negative retrieval introduces evaluation instability and underestimates model failures. To maintain consistency across skills, we enforce that no two negatives for the same caption are identical; where a caption contains fewer than K distinct primitives of type t, the same primitive may be targeted across multiple negatives subject to this constraint. This construction generalizes across all structural levels

Type	Model	Object Negatives			Attribute Negatives			Relation Negatives		
		β_O	β_A	β_R	β_O	β_A	β_R	β_O	β_A	β_R
OAR	OpenCLIP	-2.50**	+1.21**	+2.32**	+1.76**	-3.32**	-1.84	+0.60	+1.17**	-2.86**
	SigLIP 2	-2.25**	+0.65*	+1.82	+2.60**	-2.85**	-3.07	+0.70	-0.77**	-1.65*
	PE-CLIP	-2.45**	+0.96**	+1.39	+1.07	-4.13**	-0.50	+0.40	-0.09	-1.91*
	NegCLIP	-1.762**	+0.69	+0.81**	+1.51**	-3.38**	-2.45	+2.31**	+0.13	-7.13**
	CE-CLIP	-1.77**	+0.68**	+2.23*	+0.20**	-2.57**	-2.81	-1.08	+2.31**	-4.89**
	BLIP-L	-3.55**	+1.08**	+3.72	+0.50	-2.92**	+0.5	+0.2	-1.32	-2.95**
	Qwen3	-1.87**	+1.25**	+0.87	+1.95**	-3.17**	-3.71**	+0.64	-0.51	-2.66**
OA	OpenCLIP	-1.46**	-0.82*	-	+0.73**	-3.76**	-	-	-	-
	SigLIP 2	-2.27**	-1.35**		-0.01	-3.21**		-	-	-
	PE-CLIP	-1.98**	-0.52	-	+0.10	-4.00**	-	-	-	-
	NegCLIP	-2.12**	-1.08**	-	-0.08	-3.72**	-	-	-	-
	CE-CLIP	-2.62	-0.73*	-	+1.21**	-2.37**	-	-	-	-
	BLIP-L	-0.50**	+0.03	-	+0.50**	-2.53**	-	-	-	-
	Qwen3	-1.60**	+0.07	-	-0.13	-3.31**	-	-	-	-
OR	OpenCLIP	-1.45**	-	+1.89*	-	-	-	+1.05**	-	-2.42**
	SigLIP 2	-1.89**	-	+2.22**	-	-	-	+0.45	-	-1.28
	PE-CLIP	-2.03**	-	+1.49	-	-	-	+0.50	-	-1.69*
	NegCLIP	-0.95*	-	+0.65	-	-	-	+2.29**	-	-6.81**
	CE-CLIP	-2.15**	-	+3.98**	-	-	-	+1.01	-	-4.68**
	BLIP-L	-2.53**	-	+3.46**	-	-	-	-1.30**	-	+0.80
	Qwen3	-1.27**	-	+1.25	-	-	-	+0.47	-	-1.90

Table 3: **Skill Load.** Values represent change in R@1 (percentage points) per unit increase in primitive count (Eq. 4). Negative values indicate degradation, positive values indicate improvement. Shaded cells indicate the dominant source of degradation within each setting. * and ** denote statistical significance at $p < 0.005$ and $p < 0.001$, respectively. Object, attribute, and relation self-load are consistently negative across models while cross-load effects are predominantly positive.

and complexities. **Example.** As shown in Figure 2, for the caption "`There is a dark brown sofa in the living room. There is a ceiling.`", an object negative would be "`There is a dark brown chair in the living room. There is a ceiling.`", an attribute negative would be "`There is a white sofa in the living room. There is a ceiling.`" This construction yields **143K** ground-truth caption pairs for object evaluation, **63K** for attribute, and **68K** for relation, across structural levels (Table 1).

6.2 Metrics

We measure skill load by modeling how R@1, computed over skill-targeted negatives for primitive type t, varies with the counts of each primitive type. Formally:

$$\text{R@1}_t = \beta_O \, n_O(c) + \beta_A \, n_A(c) + \beta_R \, n_R(c) + \alpha \tag{4}$$

where $n_t(c)$ denotes the number of primitives of type t in caption c, β_t denotes the change in performance per unit increase in $n_t(c)$ while controlling for other types, and α denotes the intercept term. A negative coefficient β_t indicates that increasing the count of primitive type t degrades performance on skill t. Comparing β_t across primitive types allows us to distinguish self-load effects from cross-load effects for each skill. We estimate Equation 4 using OLS regression with standard errors clustered at the image level to account for within-image correlation. Statistical significance is assessed using two-sided t-tests at $p < 0.005$ (denoted *) and $p < 0.001$ (denoted **).

6.3 What Drives Skill-Level Degradation?

Table 3 reports the estimated coefficients from Eq. 4 across models and structural levels, with shaded cells indicating the dominant source of degradation with high significance within each setting. Visualization of the underlying degradation patterns across all models is provided in Appendix.

Self-load dominates cross-load. Across all models and structural levels, the largest and most consistent negative coefficients appear along the diagonal, meaning each skill degrades primarily with the count of its own primitive type. Cross-load effects are predominantly positive, driven primarily by mutual grounding between objects and attributes, as well as between objects and relations, suggesting that co-occurring primitives of different types provide useful context rather than competing for representational capacity. This pattern is consistent across model families, including standard contrastive models (OpenCLIP, SigLIPv2, PE-CLIP), models explicitly trained for compositional

reasoning (NegCLIP, CE-CLIP), and non-contrastive models (BLIP-L, Qwen3), indicating that self-load degradation is a fundamental property of current vision-language architectures rather than an artifact of any particular training objective. Among the three skills, attribute self-load is the strongest effect, consistently significant at $p < 0.001$ across all models and structural levels, indicating that attribute binding is particularly sensitive to load. Relation self-load is negative and significant across most models but smaller in magnitude, with positive cross-load from object count, suggesting that objects provide a grounding context that aids relation detection.

Discussion. These findings show that compositional degradation is not uniform across skills and cannot be reduced to a single joint reasoning bottleneck: primitive types contribute unequally to performance degradation, and treating caption complexity as a single scalar obscures these distinctions. Performance degradation is largely decomposable into skill-specific self-load effects, with cross-load effects providing modest grounding benefits rather than additional interference. Notably, NegCLIP and CE-CLIP, which reverse the compositional integration gap (Section 5.3), still exhibit consistent self-load degradation across all structural levels, suggesting that compositional training objectives address joint reasoning difficulty but leave the underlying load sensitivity of individual skills unresolved.

7 Conclusion

Despite extensive efforts to benchmark and improve compositional understanding in vision-language models, the sources of performance degradation remain poorly understood. Through COMPASS, a controlled evaluation framework that constructs captions from scene graphs with explicit object, attribute, and relation structure, we show that the observed degradation with increasing caption complexity reflects multiple separable factors. Compositional integration introduces a measurable cost, yet accounts for only part of the observed degradation; even models that handle joint reasoning well continue to suffer as caption complexity grows. Skill load analysis reveals a more fundamental pattern: each skill degrades primarily under the weight of its own primitive count, with cross-load effects providing grounding benefits rather than additional interference, and this holds consistently across standard, compositional, and non-contrastive model families. These findings reframe compositional failure as a multi-factorial problem in which joint reasoning accounts for only part of the observed degradation. Progress on compositional robustness may therefore require skill-specific interventions that ensure individual skills remain robust as the primitive load increases, rather than focusing solely on joint reasoning mechanisms. COMPASS provides a controlled structure for diagnosing these factors independently and tracking progress along each dimension.

Limitations and Future Work. Several limitations point to directions for future work. COMPASS is constructed from Visual Genome scene graphs using synthetically generated captions, following a long line of work that uses controlled synthetic data for compositional evaluation (Ma et al., 2023; Johnson et al., 2016), which may not fully reflect the distribution of naturally occurring descriptive language, and inherits known annotation biases toward certain object and relation types; future work could extend COMPASS to naturally occurring captions and more diverse scene graph sources. Furthermore, because objects provide the necessary grounding context for both attributes and relations, our framework evaluates attribute and relation skills in the presence of objects rather than in complete isolation, which is an inherent constraint of scene-graph-based evaluation. Developing evaluation protocols that further disentangle these dependencies remains an open challenge. Finally, our retrieval-based evaluation protocol, which is standard in prior compositionality benchmarks (Hsieh et al., 2023; Ma et al., 2023; Yuksekgonul et al., 2023), isolates discrimination ability but does not extend to generative VLM settings, where failure modes may differ; adapting COMPASS's controlled-complexity structure to generative evaluation is a natural direction for future work. Beyond these methodological limitations, our findings also raise deeper questions about the origin of self-load degradation. The pattern in which each skill degrades specifically under the weight of its own primitive count, while cross-load effects are positive, is suggestive of a representational account: models may have limited capacity to maintain robust encodings of multiple instances of the same primitive type, with degradation reflecting competition within rather than across primitive types. Whether this reflects encoding failures, cross-modal alignment failures, or both remains an open question for future work.

References

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

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv*, 2025.

Leon Bottou. From machine learning to machine reasoning, 2011. URL https://arxiv.org/abs/1102.1808.

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

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

Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision&language concepts to vision&language models, 2023b. URL https://arxiv.org/abs/2211.11733.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding, 2021. URL https://arxiv.org/abs/2106.09141.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: how do neural networks generalise?, 2020. URL https://arxiv.org/abs/1908.08351.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

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, 2024. URL https://arxiv.org/abs/2211.13854.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

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

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. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016. URL http://arxiv.org/abs/1602.07332.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022a.

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

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *arXiv preprint arXiv:2212.07796*, 2023.

OpenAI. Gpt-4o system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

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

Barbara H. Partee. Compositionality. In Fred Landman and Frank Veltman (eds.), *Varieties of Formal Semantics*, pp. 281–311. Foris, Dordrecht, 1984.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL https://arxiv.org/abs/2306.14824.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

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.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL http://arxiv.org/abs/1908.10084.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 255–265, 2017.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Vishaal Udandarao, Mehdi Cherti, Shyamgopal Karthik, Jenia Jitsev, Samuel Albanie, and Matthias Bethge. A good crepe needs more than just sugar: Investigating biases in compositional vision-language benchmarks, 2025. URL https://arxiv.org/abs/2506.08227.

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

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

Qi Zhang, Yuxu Chen, Lei Deng, and Lili Shen. Abe-clip: Training-free attribute binding enhancement for compositional image-text matching, 2025. URL https://arxiv.org/abs/2512.17178.

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

Xiangru Zhu, Penglei Sun, Chengyu Wang, Jingping Liu, Zhixu Li, Yanghua Xiao, and Jun Huang. A contrastive compositional benchmark for text-to-image synthesis: A study with unified text-to-image fidelity metrics, 2023. URL https://arxiv.org/abs/2312.02338.

A Details About Hard Negative Generation

Figure 4 shows the GPT-4o mini prompts used for caption and negative generation, as described in Sections 3.1 and 3.2, respectively. For caption generation (A), we use a few-shot prompt that instructs the model to convert scene graph primitives into fluent natural language sentences, without introducing new objects or relations beyond those specified. For hard negative generation (B-D), we use separate few-shot prompts for each primitive type. Each prompt instructs the model to generate semantically plausible alternatives that are neither synonyms nor identical to the original primitive, with a preference for opposites where possible.

Candidate caching. To avoid redundant API calls, generated candidates are stored in a per-primitive-type dictionary mapping each primitive to its list of alternatives. When a negative is needed for a given primitive, we first check the dictionary; if no valid candidate exists, we query GPT-4o mini and append the results to the dictionary. This allows candidates to be reused across captions that share the same primitives.

Filtering. To construct a hard negative, we replace the target primitive with a candidate alternative and adjust the caption for grammatical correctness using Language Tool[1] before any filtering is applied. The resulting caption is then evaluated using Sentence Transformers (Reimers & Gurevych, 2019) to ensure semantic distinctiveness from the original caption. A candidate is discarded if its similarity score to the original caption exceeds a threshold: we use 0.9 for captions with complexity $N \geq 6$ and 0.95 for $N < 6$. At lower complexities, a single word change has a larger impact on overall caption similarity, so we apply a stricter threshold to ensure candidates are sufficiently distinct. At higher complexities, the same substitution contributes less to the overall similarity score, requiring a more relaxed threshold to avoid discarding valid candidates. This filtering is applied consistently across composed, decomposed, and skill-targeted negatives.

Primitive selection for skill-targeted negatives. For skill-targeted negatives, primitives are selected using a weighted sampling scheme. Initially all primitives of the target type are assigned equal weights; once a primitive is selected, its weight is halved to encourage diversity across the $K = 4$ negatives per caption. We make up to 10 attempts to generate 4 valid negatives per caption; captions for which 4 valid negatives cannot be obtained are discarded from evaluation.

Composed and decomposed negatives. For composed and decomposed settings, each negative must target exactly one primitive. Once a valid negative is obtained for a given primitive, that primitive is not used again for the same caption. We make up to 3 attempts to obtain a valid negative per primitive; if a valid negative cannot be found for all primitives in a caption, that instance is discarded from evaluation.

[1] https://pypi.org/project/language-tool-python/

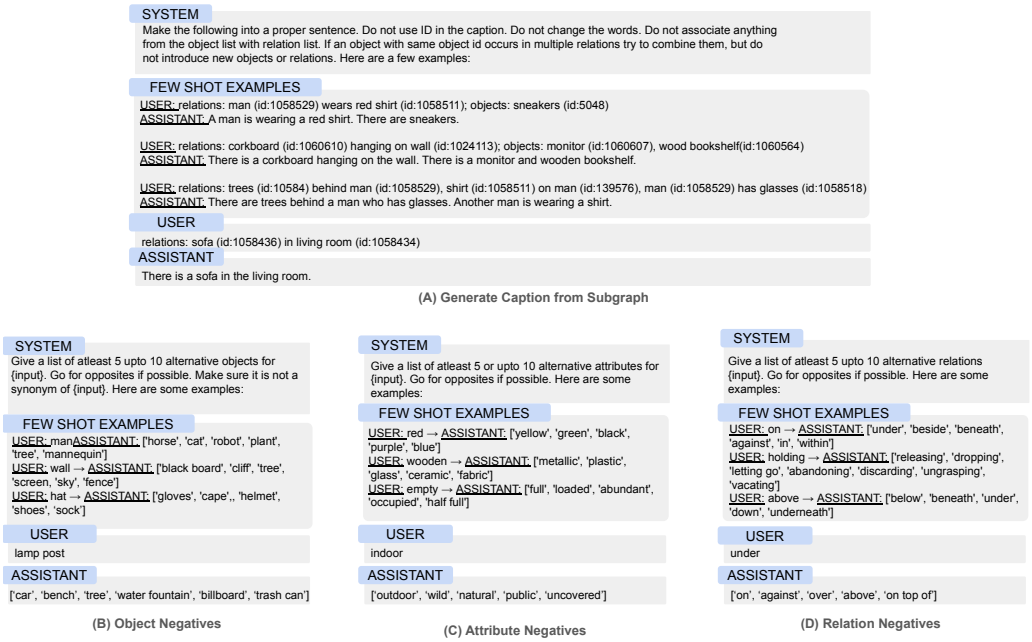


Figure 4: GPT-4o mini prompts used for caption and negative generation. (A) Few-shot prompt for generating natural language captions from scene subgraphs, instructing the model to combine primitives into fluent sentences without introducing new objects or relations. (B-D) Few-shot prompts for generating hard negative candidates for object (B), attribute (C), and relation (D) primitives respectively, eliciting semantically plausible alternatives that are neither synonyms nor identical to the original primitive.

B Perplexity Analysis

B.1 Overview

To validate that COMPASS hard negatives cannot be distinguished from ground-truth captions using linguistic statistics alone, we conduct a systematic perplexity audit. Motivated by SugarCREPE Hsieh et al. (2023), which identifies perplexity-based exploitability as a critical flaw in substitution-based benchmarks, we measure the distributional difference between ground-truth and negative caption perplexities across all primitive types and structural levels. We compute token-level perplexity under GPT-2 Radford et al. (2019) for all ground-truth captions and their corresponding hard negatives. A benchmark where negatives are systematically less fluent than ground-truth captions is exploitable by text-only models, as a model can identify the ground-truth caption by selecting the lowest-perplexity candidate without any visual input.

B.2 Evaluating on the Hardest Negative

For each ground-truth caption with $K = 4$ skill-targeted hard negatives, we identify the *hardest negative* — the candidate whose perplexity is closest to the ground-truth caption:

$$\hat{n} = \arg\min_{n \in \mathcal{N}} |\text{PPL}(n) - \text{PPL}(c)| \tag{5}$$

This choice is motivated by the structure of our retrieval task. In $K = 4$ evaluation, a model achieves R@1$= 1$ only if the ground-truth caption scores higher than *all four* negatives. For a text-only model exploiting perplexity to succeed on a trial, it must identify the ground-truth as having lower perplexity than all four negatives — including the one most similar in perplexity to the ground truth. If even this hardest-to-exploit negative is linguistically indistinguishable from the ground truth, then the full $K = 4$ task cannot be reliably solved via perplexity alone. Measuring artifact severity on the hardest

negative therefore represents a *worst-case bound* on exploitability: if the effect size is negligible on this subset, it is negligible across the entire evaluation.

B.3 Effect Size Metric

$$r = 1 - \frac{2U}{n_1 n_2} \tag{6}$$

where U is the Mann-Whitney statistic and n_1, n_2 are the sample sizes of the two groups. The rank-biserial correlation measures the probability that a randomly drawn ground-truth perplexity exceeds a randomly drawn negative perplexity, expressed as a deviation from chance. We interpret effect sizes following standard conventions: $r < 0.1$ indicates a negligible difference, $0.1 \leq r < 0.3$ indicates a small difference, and $r \geq 0.3$ indicates a medium or large difference. We use this metric rather than a text-only retrieval baseline because it directly characterizes the distributional overlap between ground-truth and negative perplexities, without dependence on task format or chance level definitions that vary across evaluation settings.

B.4 Results: Skill-Targeted Negatives

Table **??** reports effect sizes for skill-targeted negatives across all structural levels, measured on the hardest negative subset.

Level	Negative Type	Effect size r	Interpretation
L3 (OAR)	Object	0.096	Negligible
L3 (OAR)	Attribute	0.072	Negligible
L3 (OAR)	Relation	0.259	Small
L2 (OA)	Object	0.062	Negligible
L2 (OA)	Attribute	0.099	Negligible
L2 (OR)	Object	0.099	Negligible
L2 (OR)	Relation	0.267	Small

Table 4: Perplexity effect sizes (rank-biserial r) measuring the distributional difference between ground-truth and negative caption perplexities for skill-targeted negatives, evaluated on the hardest negative subset.

Object and attribute negatives exhibit negligible effect sizes across all structural levels ($r = 0.062$–0.099), confirming that ground-truth and negative captions are linguistically indistinguishable under worst-case selection. Relation negatives show slightly larger but still small effect sizes ($r = 0.259$–0.267), consistent with stronger collocational constraints on spatial prepositions compared to content words. Across all primitive types and structural levels, effect sizes remain within the small range by standard conventions, validating the skill load results reported in Section 6.

B.5 Results: Compositional Integration Negatives

Table **??** reports effect sizes for composed captions across structural levels, measured on the hardest negative subset.

Level	Effect size r	Interpretation
L3 (OAR)	0.060	Negligible
L2 (OA)	0.152	Small
L2 (OR)	0.071	Negligible

Table 5: Perplexity effect sizes (rank-biserial r) measuring the distributional difference between ground-truth and negative caption perplexities for composed captions, evaluated on the hardest negative subset.

Composed captions exhibit negligible to small perplexity artifacts across all structural levels ($r = 0.060$–0.152), confirming that the composed evaluation setting is not systematically exploitable via linguistic statistics. This validates the composed side of the compositional integration gap analysis reported in Section 5.

Decomposed perplexity results. Decomposed captions are structurally short by design - each targets one primitive, yielding captions of typically 6–8 tokens (e.g., "There is a sofa in the image."). At this length, substituting a single word constitutes a proportionally large fraction of the caption's total token sequence, and perplexity is dominated by the corpus frequency of the substituted word rather than any meaningful fluency difference. These artifacts are a structural consequence of caption length rather than the negative generation pipeline, and decomposed perplexity results are therefore not informative for benchmark validity. The composed setting, which forms one side of the integration gap computation, is artifact-free across all structural levels ($r = 0.060$–0.152).

C Compositional Integration Gap Raw Results

The compositional integration gap Δ measures the difference between decomposed and composed accuracy, quantifying the additional cost introduced by joint reasoning over multiple primitives. A positive Δ indicates that models perform better when primitives are evaluated independently than when they must be resolved jointly. We report per-complexity results across all structural levels here; aggregate results are discussed in Section 5.3.

	OpenCLIP			SigLIP 2			PE-CLIP			NegCLIP			CE-CLIP			BLIP-L			Qwen3		
Complexity Level	**C**	**D**	**Δ**	**C**	**D**	**Δ**	**C**	**D**	**Δ**	**C**	**D**	**Δ**	**C**	**D**	**Δ**	**C**	**D**	**Δ**	**C**	**D**	**Δ**
2	55.69	55.26	-0.42	58.70	65.66	6.96	65.48	62.20	-3.28	62.08	56.61	-5.47	62.90	56.14	-6.75	63.51	64.37	0.85	50.38	63.87	13.49
3	43.42	47.55	4.13	41.33	56.19	14.86	54.79	54.23	-0.56	50.40	46.69	-3.71	57.98	42.51	-15.46	47.94	53.14	5.19	34.7	53.43	18.72
4	35.13	37.01	1.87	34.04	47.78	13.75	45.61	44.12	-1.49	39.57	35.47	-4.11	51.76	31.58	-20.17	39.86	44.89	5.02	26.43	43.29	16.86
5	29.61	31.77	2.16	28.77	44.00	15.23	38.23	39.83	1.60	33.12	31.37	-1.75	48.68	26.08	-22.60	34.47	36.72	2.25	19.71	36.64	16.93
6	22.71	25.51	2.79	21.94	36.91	14.96	30.64	31.95	1.31	28.75	25.18	-3.57	44.18	20.54	-23.64	29.01	30.46	1.44	17.29	31.45	14.16
7	18.82	21.00	2.17	18.61	32.32	13.71	26.32	27.29	0.97	23.25	19.93	-3.32	42.01	16.65	-25.35	25.17	26.17	1.00	14.6	26.97	12.37
8	15.33	19.00	3.67	16.49	30.27	13.78	22.45	25.28	2.83	19.28	15.91	-3.37	36.45	13.11	-23.33	22.45	21.83	-0.61	12.3	23.58	11.27
9	13.01	14.55	1.53	12.56	26.04	13.47	18.57	21.77	3.19	16.75	13.97	-2.78	33.47	10.50	-22.96	20.06	19.44	-0.62	10.66	20.76	10.11
10	10.64	13.72	3.07	11.39	22.03	10.64	16.40	18.69	2.29	11.39	11.43	0.04	29.37	8.05	-21.32	16.44	16.27	-0.17	10.17	18.57	8.40
11	8.03	11.29	3.25	9.19	19.71	10.53	14.35	16.08	1.72	11.82	9.23	-2.58	28.01	6.96	-21.05	14.44	14.16	-0.28	9.68	13.45	3.77
12	7.46	9.59	2.12	7.93	19.23	11.30	11.77	13.53	1.76	8.09	8.97	0.88	26.62	6.24	-20.38	12.64	11.66	-0.98	8.40	13.62	5.22
mean	-	-	2.39	-	-	12.65	-	-	0.94	-	-	-2.70	-	-	-20.27	-	-	1.19	-	-	11.94
std dev	-	-	1.23	-	-	2.53	-	-	1.96	-	-	1.83	-	-	5.17	-	-	2.17	-	-	4.81

Table 6: **Compositional Integration L2 Object-Attribute Accuracy (Raw Results).** We report the difference between decomposed (D) and composed (C) accuracy per complexity level, along with the mean and standard deviation of Δ for each model in OA setting. Positive Δ values indicate that model performs better independently than jointly, highlighting compositional binding difficulty.

The results in Table 6 indicate a clear integration gap with positive Δ values for most models, where the additional cost of the joint compositional reasoning consistently dominates the pure cost of the independent primitive recognition. In contrast, both NegCLIP and CE-CLIP serve as notable exceptions to that observation. Both models consistently display negative Δ values, which is a clear indicator of superior performance on composed tasks compared to their decomposed counterparts. This behavior is more strongly demonstrated in CE-CLIP, with a relatively low mean of -20.27 forΔ. These two exceptional cases might imply that through specialized training objectives such as compositional binding or contrastive discrimination of hard negatives, the model's behavior in composed and decomposed settings can be altered.

The Object-Relation results in 7 follow the trend of considerable integration penalty across different models, with the most severe performance drops in OpenCLIP and SigLIPv2 models. Remarkably, in this context, NegCLIP does not retain the outlier position it has in the Object-Attribute setting, shifting to positive Δ values across all complexity levels. Conversely, CE-CLIP continues to be a counterpoint in this setting as well, with a negative mean Δ of -4.96. This suggests that CE-CLIP's contrastive intra-modal and cross-modal training objectives might effectively mitigate the primary integration burden in joint compositional reasoning.

In most structurally complex scene composition (OAR) settings, the results in Table 8 show a considerable integration overhead, despite lower discrepancies between composed and decomposed settings with lower Δ than in simpler OR and OA settings. Interestingly, NegCLIP showed the

Complexity Level	OpenCLIP			SigLIP 2			PE-CLIP			NegCLIP			CE-CLIP			BLIP-L			Qwen3		
	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ
3	28.79	41.34	12.54	18.89	33.73	14.84	34.71	36.30	1.59	41.46	46.17	4.71	49.90	38.49	-11.41	44.23	39.55	-4.69	33.11	45.4	12.29
4	21.39	37.81	16.42	12.71	28.69	15.99	28.00	30.60	2.61	35.01	39.12	4.11	41.24	31.86	-9.38	29.07	34.04	4.97	24.61	40.46	15.84
5	17.80	34.75	16.95	9.21	24.66	15.46	22.85	26.94	4.09	27.79	33.14	5.35	33.55	23.58	-9.96	22.68	28.70	6.02	20.68	33.55	12.87
6	10.13	28.03	17.89	5.85	14.68	8.83	14.53	18.07	3.53	19.63	25.19	5.56	23.89	18.36	-5.53	16.97	19.77	2.80	16.86	24.46	7.59
7	7.50	26.42	18.92	4.76	13.16	8.40	13.10	16.02	2.92	15.12	21.87	6.75	19.94	15.62	-4.32	13.22	17.58	4.36	13.56	22.98	9.41
8	5.93	23.93	18.01	4.20	10.26	6.06	9.44	11.89	2.44	10.75	19.11	8.37	15.17	12.50	-2.67	11.04	13.74	2.70	12.94	18.83	5.88
9	4.31	22.68	18.37	2.65	8.40	5.75	6.71	9.17	2.46	8.69	15.21	6.52	11.98	11.05	-0.93	8.72	10.70	1.98	10.91	14.72	3.81
10	4.17	22.46	18.29	2.40	7.31	4.91	6.24	8.44	2.21	6.68	12.88	6.21	10.58	8.51	-2.07	7.01	9.35	2.34	10.45	13.15	2.69
11	2.30	21.01	18.72	2.16	5.31	3.15	4.29	6.99	2.71	6.96	10.87	3.91	8.36	7.37	-0.99	6.68	7.54	0.86	9.38	10.97	1.59
12	2.45	21.94	19.49	1.54	4.63	3.09	2.91	5.26	2.35	5.78	9.78	4.00	8.27	5.89	-2.38	4.77	6.41	1.65	8.33	9.46	1.13
mean	-	-	17.56	-	-	8.64	-	-	2.69	-	-	5.54	-	-	-4.96	-	-	2.30	-	-	7.31
std dev	-	-	1.98	-	-	5.04	-	-	0.70	-	-	1.44	-	-	3.93	-	-	2.93	-	-	5.16

Table 7: **Compositional Integration L2 Object-Relation Accuracy (Raw Results).** We report the difference between decomposed (D) and composed (C) accuracy per complexity level, along with the mean and standard deviation of Δ for each model in OR setting. Positive Δ values indicate that model performs better independently than jointly, highlighting compositional binding difficulty.

Complexity Level	OpenCLIP			SigLIP 2			PE-CLIP			NegCLIP			CE-CLIP			BLIP-L			Qwen3		
	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ	C	D	Δ
4	20.54	19.85	-0.69	18.99	25.48	6.48	25.56	28.06	2.49	33.78	36.66	2.88	41.03	31.15	-9.88	33.47	33.08	-0.38	25.1	35.98	10.87
5	16.83	18.34	1.51	17.60	22.05	4.44	23.79	26.31	2.52	27.14	29.94	2.79	34.80	23.10	-11.69	24.85	27.14	2.29	19.44	31.51	12.06
6	10.53	12.31	1.78	12.98	15.63	2.64	16.69	18.80	2.11	20.77	24.86	4.08	25.06	17.21	-7.84	19.19	20.68	1.49	16.21	23.96	7.74
7	7.72	8.59	0.86	9.11	11.46	2.35	11.51	13.29	1.77	15.06	19.52	4.46	20.10	15.45	-4.65	14.58	16.69	2.11	14.21	21.36	7.15
8	5.73	6.52	0.79	6.71	8.69	1.97	9.04	11.70	2.66	12.05	16.05	4.00	16.79	12.64	-4.15	11.16	14.03	2.86	12.7	17.79	5.09
9	4.51	4.97	0.46	5.69	6.73	1.03	6.78	9.01	2.22	8.64	13.10	4.45	13.77	8.49	-5.28	9.32	9.63	0.31	10.67	13.57	2.9
10	2.92	3.80	0.89	3.96	6.20	2.24	5.84	7.04	1.19	7.35	10.53	3.18	11.50	7.92	-3.59	6.99	8.81	1.82	10.33	10.54	0.2
11	2.36	3.11	0.75	3.21	5.03	1.82	4.28	6.05	1.76	5.08	9.05	3.96	8.24	6.80	-1.44	6.48	7.12	0.64	7.92	9.9	1.98
12	2.07	3.15	1.08	2.99	3.53	0.54	4.35	5.32	0.97	4.45	8.37	3.91	7.39	5.00	-2.39	5.05	6.36	1.30	8.59	8.64	0.05
mean	-	-	0.83	-	-	2.61	-	-	1.97	-	-	3.75	-	-	-5.66	-	-	1.38	-	-	5.34
std dev	-	-	0.70	-	-	1.82	-	-	0.59	-	-	0.64	-	-	3.45	-	-	1.03	-	-	4.42

Table 8: **Compositional Integration L3 Object-Attribute-Relation (Raw Results).** We report the difference between decomposed (D) and composed (C) accuracy per complexity level, along with the mean and standard deviation of Δ for each model in OAR setting. Positive Δ values indicate that model performs better independently than jointly, highlighting compositional binding difficulty.

largest gap in this setting, with a mean Δ of 3.75, reinforcing the observations in the OR and OA settings that hard-negative discrimination with attributes does not directly translate to whole-scene compositions. In contrast, CE-CLIP maintains its position as a robust outlier, with a much larger gap in the opposite direction, a mean Δ of -5.66. This further signifies CE-CLIP's effectiveness in reinforcing compositional integration.

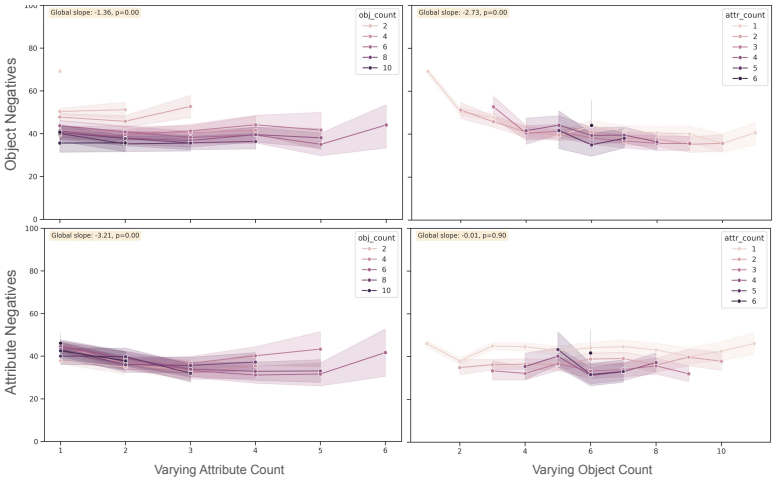
D Compute Resources

Caption generation and hard negative generation were performed using the GPT-4o-mini API OpenAI (2024). All model evaluations and perplexity computations were run on a single NVIDIA A100 GPU. Post-processing and regression analysis were performed on CPU with 4 parallel workers. Among the evaluated models, all contrastive models (OpenCLIP, SigLIPv2, NegCLIP, PE-CLIP) and BLIP completed evaluation within 2–3 hours per model. Qwen3-VL-Embedding-8B required approximately 8 hours due to its larger model size.

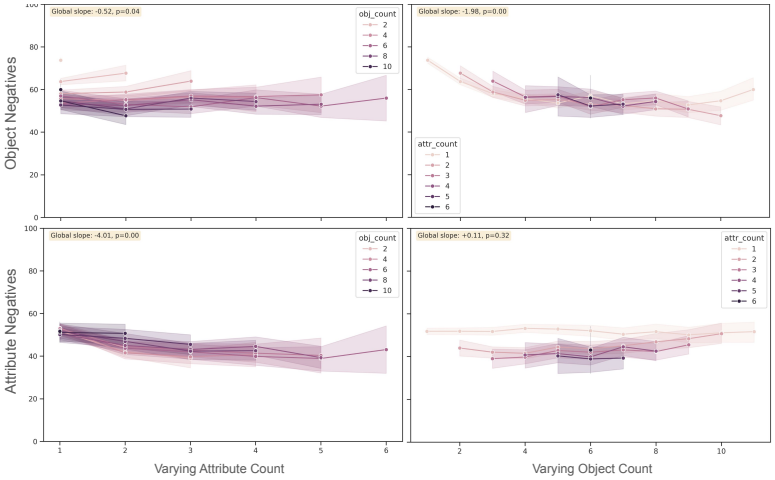
E Skill Load: Visualization

To complement the regression coefficients reported in Table 3, we visualize the underlying degradation patterns directly. For each model, we plot R@1 accuracy as a function of one primitive count while holding the other fixed, with separate lines for each value of the controlled count. Each figure shows four panels per model: object negatives (top) and attribute negatives (bottom), each varying attribute count (left) and object count (right). The global slope reported in each panel corresponds to the regression coefficient in Table 3, estimated via OLS while controlling for all primitive counts.

Figures 5–7 show results for L2 (OA) captions across all seven models. The pattern is consistent throughout: for object negatives, varying the attribute count produces flat lines while varying the object count produces consistently declining lines. For attribute negatives, varying the object count



(a) SigLIPv2



(b) PE-CLIP

Figure 5: Skill load analysis on L2 (OA) captions for SigLIPv2 and PE-CLIP. Each model shows four panels: object negatives (top) and attribute negatives (bottom), each varying attribute count (left) and object count (right) while holding the other fixed.

produces flat or inconsistent lines while varying the attribute count produces steep and consistent declines. This directly illustrates the self-load dominance reported in Table 3, each skill degrades primarily under the weight of its own primitive count, with cross-load effects remaining small and inconsistent across all models and architectures.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the paper introduces COMPASS, a controlled evaluation framework for diagnosing compositional failure in VLMs. The claims: compositional integration gap accounts for only part of observed degradation, and self-load dominates cross-load are directly supported by the experimental results in Tables 2, 3, and the CREPE analysis in Figure 3.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a dedicated 'Limitations and Future Work' section at the end of the Conclusion (Section 7). It acknowledges that: (1) COMPASS relies on synthetically generated captions from Visual Genome scene graphs, which may not reflect natural language distributions and inherit annotation biases; (2) attribute and relation skills are always evaluated in the presence of objects due to the grounding requirement of scene-graph-based evaluation; and (3) the retrieval-based evaluation protocol does not extend to generative VLM settings.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not include theoretical results or formal proofs.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper describes the full dataset construction pipeline in Section 3, including the scene graph sampling procedure, caption generation using GPT-4o mini with few-shot prompts (shown in Figure 4 and the Appendix), hard negative generation methodology, and the retrieval-based evaluation protocol. The models evaluated are publicly available and cited. The evaluation metrics are formally defined in Sections 5.2 and 6.2. Dataset statistics are provided in Table 1.

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

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

   Answer: [Yes]

   Justification: The COMPASS dataset is publicly available on HuggingFace (`https://huggingface.co/datasets/Anon-compass/COMPASS`). Code is provided anonymously at `https://anonymous.4open.science/r/skill-comp-B276/`.

6. **Experimental setting/details**

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

   Answer: [Yes]

Justification: Section 4 describes the experimental setup in detail, including all seven VLMs evaluated (OpenCLIP, SigLIPv2, PE-CLIP, NegCLIP, CE-CLIP, BLIP-L, Qwen3-VL-Embedding-8B) with citations. The unified retrieval protocol using similarity-based ranking is described. Section 3.2 details hard negative generation. This is an evaluation-only paper with no training of new models, so hyperparameters and optimizers are not applicable

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Table 2 reports mean ± standard deviation of the Compositional Integration Gap (Δ) across complexity levels for all models. Raw per-complexity results with individual Δ values are provided in Tables 6, 7 and 8 in the Appendix, enabling full transparency of variance across conditions. Table 3 (Skill Load) reports statistical significance levels for all regression coefficients, with * denoting $p < 0.005$ and ** denoting $p < 0.001$.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix D notes these details.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The paper presents a controlled evaluation framework for vision-language models using the Visual Genome dataset, a publicly available academic resource. No sensitive data was collected, and the research does not enable direct harmful applications. The work aims to improve understanding of model limitations, which is a broadly beneficial research goal.

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The paper presents a diagnostic evaluation benchmark for compositional reasoning in VLMs. It does not introduce a new model, training method, or deployment system. As foundational evaluation research, it does not have a direct path to negative societal applications. The positive impact is enabling more precise diagnosis of model failures, which can inform development of more reliable vision-language systems.

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release a generative model, pre-trained language model, image generator, or scraped dataset with high misuse risk. COMPASS is an evaluation benchmark derived from the existing Visual Genome dataset using scene-graph-based caption construction. No novel high-risk assets are released.

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper properly cites all datasets, models, and tools used. Visual Genome (Krishna et al., 2016) is cited as the source of scene graphs. All evaluated models: OpenCLIP, SigLIPv2, PE-CLIP, NegCLIP, CE-CLIP, BLIP-L, and Qwen3-VL-Embedding, are cited with their original papers. GPT-4o mini (OpenAI, 2024) is cited for caption and negative generation. Sentence Transformers (Reimers & Gurevych, 2019) are cited for filtering.

Guidelines:

13. **New assets**

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

Answer: [Yes]

Justification: The paper introduces COMPASS, a new evaluation benchmark. Its construction is thoroughly documented in Section 3, including the scene graph sampling procedure, structural level definitions (L1/L2/L3), caption generation pipeline with GPT-4o mini prompts (Figure 4), hard negative construction methodology, and dataset statistics (Table 1). Upon release, full documentation and anonymized assets will be provided.

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. All data generation is automated using GPT-4o mini for caption and negative synthesis, and all evaluation is performed programmatically using pre-trained VLMs.

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

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

Answer: [N/A]

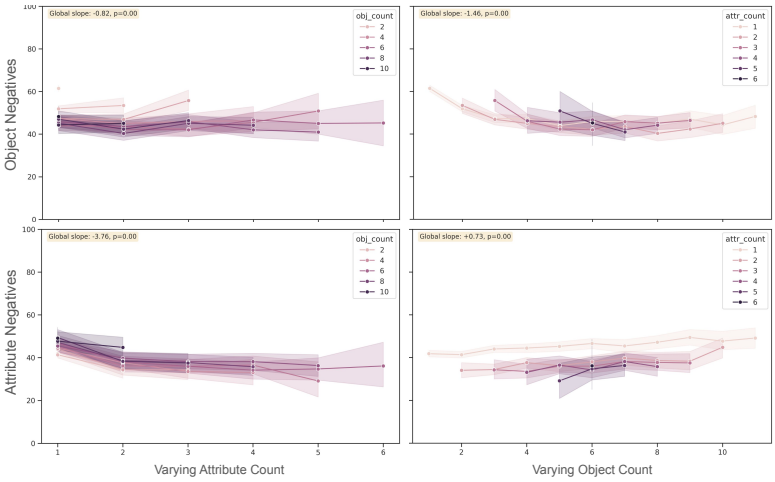
Justification: The paper does not involve crowdsourcing or research with human subjects. No IRB approval is required.

16. **Declaration of LLM usage**

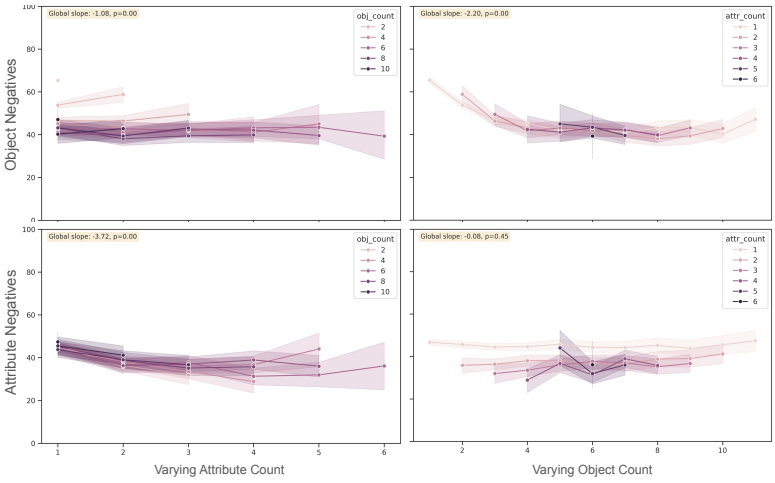
Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality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

Answer: [Yes]

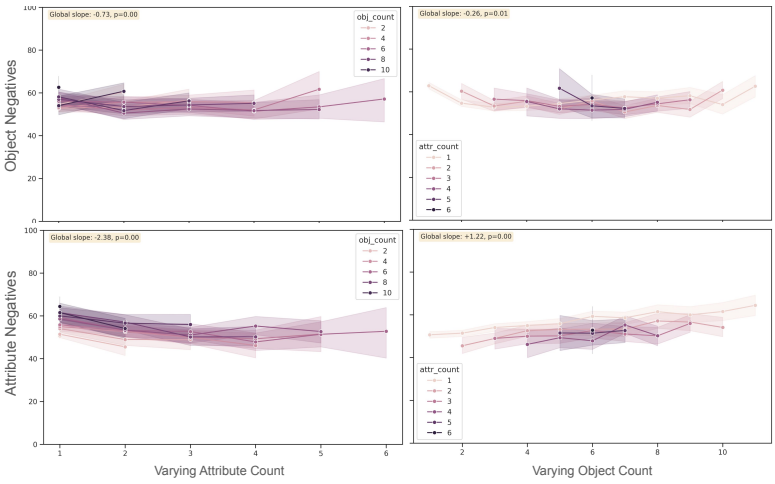
Justification: The paper explicitly describes and documents its use of GPT-4o mini (OpenAI, 2024) as a core component of the COMPASS data construction pipeline. GPT-4o mini is used for two tasks: (1) generating natural language captions from structured scene graph primitives, and (2) generating hard negative candidates for objects, attributes, and relations. The few-shot prompts used for both tasks are provided in Figure 4 and the Appendix.



(a) OpenCLIP

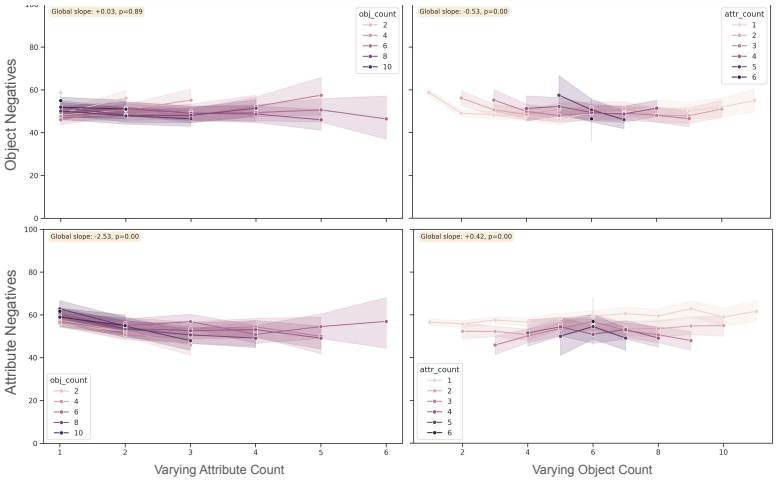


(b) NegCLIP

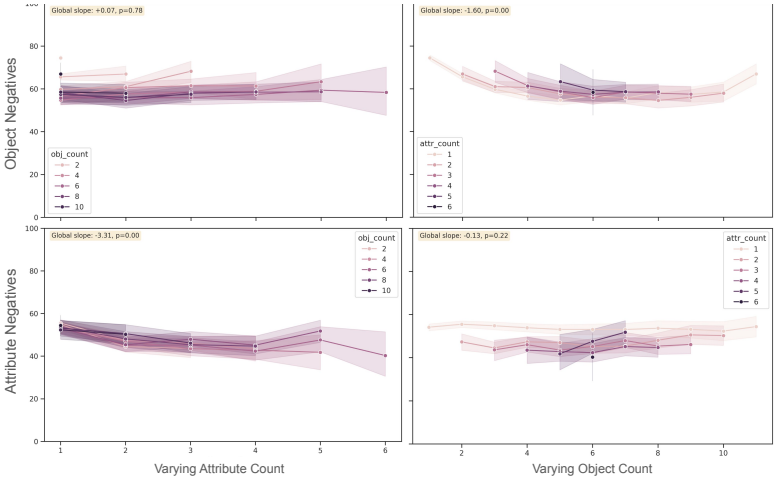


(c) CE-CLIP

Figure 6: Skill load analysis on L2 (OA) captions for OpenCLIP, NegCLIP, and CE-CLIP. Layout follows Figure 5.



(a) BLIP-L



(b) Qwen3-VL-Embedding

Figure 7: Skill load analysis on L2 (OA) captions for BLIP and Qwen3-VL-Embedding. Layout follows Figure 5.