Long Story Short: Disentangling Compositionality and Long-Caption Understanding in Contrastive VLMs

Israfel Salazar[1] **Desmond Elliott**[1] **Yova Kementchedjhieva**[2]
[1]Department of Computer Science, University of Copenhagen [2]MBZUAI
{israfel.salazar, de}@di.ku.dk
yova.kementchedjhieva@mbzuai.ac.ae

Abstract

Contrastive vision-language models (VLMs) have made significant progress in binding visual and textual information, yet understanding long, compositional captions remains an open challenge. While these capabilities are often assumed to be closely related, the conditions under which they reinforce each other remain unclear. In this paper, we empirically analyze when compositional reasoning and long-caption understanding transfer across tasks, and when this relationship fails. Through controlled experiments across diverse training objectives, datasets, and architectural designs, we find a bidirectional but sensitive relationship between the two capabilities. Models trained on poorly grounded captions or with limited parameter updates fail to generalize, while high-quality long-caption data with strong visual grounding promotes both capabilities simultaneously. We further show that architectural choices aimed at preserving general alignment, such as frozen positional embeddings, can inadvertently limit compositional learning. Our analysis provides actionable guidelines for data selection and model design to improve VLM generalization.

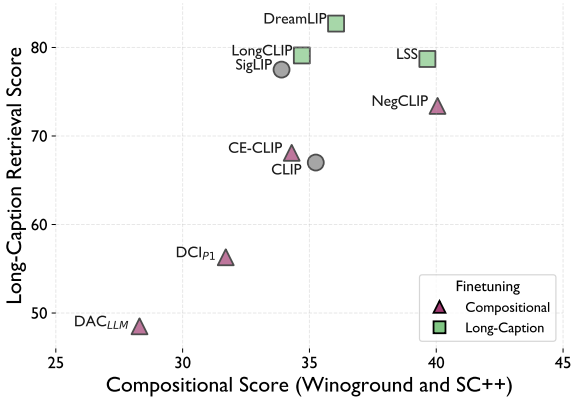

Figure 1: **Relationship between compositional reasoning and long-caption retrieval performance.** Models show a positive correlation between the two capabilities, but this relationship is sensitive to training setups, where architectural or optimization choices can limit compositional generalization.

1 Introduction

Understanding real-world images goes beyond the recognition of objects; it requires reasoning about their attributes and relationships within a scene. Captions that comprehensively describe such scenes are typically long, conveying not just more information but also greater compositional complexity. While vision-language models (VLMs) have made progress in understanding the relationship between images and text (Radford et al., 2021; Jia et al., 2021; Chen et al., 2024a), their ability to interpret long, dense captions remains limited (Yamada et al., 2022; Kamath et al., 2023; Thrush et al., 2022; Garg et al., 2024).

Despite growing interest in both compositionality and long-caption understanding, current multimodal benchmarks typically treat these capabilities in isolation. Compositionality has long been a focus of neural networks research (Lake and Baroni, 2018; Bahdanau et al., 2019; Hupkes et al., 2020), but it has received limited attention in the vision-language literature (Nikolaus et al., 2019; Surís et al., 2020). More recently, several benchmarks have emerged to evaluate this ability in VLMs (Yuksekgonul et al., 2022; Thrush et al., 2022; Zhao et al., 2022; Hsieh et al., 2023; Dumpala et al., 2024). However, these datasets often rely on short, generic captions and primarily test symbolic or relational binding, overlooking the linguistic and visual complexity present in real-world scenes. Meanwhile, benchmarks targeting long-caption understanding often use longer, dense captions (Cho et al., 2022; Zhang et al., 2024a; Urbanek et al., 2024; Onoe et al., 2024; Garg et al., 2024), but are rarely analyzed through the lens of compositionality (Urbanek et al., 2024). As a result, the interplay between compositionality and long-caption understanding remains underexplored.

Findings of the Association for Computational Linguistics: ACL 2026, pages 42994–43008

Although compositionality is commonly viewed as both a prerequisite for and a consequence of long-caption understanding, models trained for long captions do not consistently exhibit strong compositional reasoning, and compositional models do not always generalize to long captions. We therefore ask two questions: (1) Does compositional training improve long-caption understanding? and (2) Does training on long, dense captions foster compositional generalization? To address these questions, we conduct a comparative analysis of models trained with distinct objectives, some targeting compositionality, others optimized for processing long captions. We evaluate them on a suite of benchmarks spanning compositional reasoning, long-caption retrieval (as a proxy for understanding dense descriptions), and general vision-language alignment, allowing us to disentangle the contribution of each skill and to assess whether gains transfer across tasks or remain narrowly scoped.

Our findings reveal a bidirectional relationship between compositionality and long-caption understanding, shaped by several important factors. Gains are sensitive to data quality: models trained on poorly structured or weakly grounded captions fail to generalize, while datasets with dense, diverse captions and broad vocabulary coverage lead to stronger performance across both tasks. We further identify trade-offs: Both compositional and long-caption training can degrade performance on general vision-language benchmarks, likely due to distributional shift. Likewise, training decisions that preserve general alignment, such as limited parameter updates or aggressive regularization, tend to underperform even when trained on high-quality data. We observe that transfer between compositional reasoning and long-caption understanding is highly sensitive to training design. In several cases, targeted improvements fail to transfer, and models exhibit structured trade-offs rather than uniform gains. However, even with these challenges, we show that models trained on well-designed long-caption data can achieve strong results in both long-caption retrieval and compositional reasoning.

2 Background

Recent work has proposed architectural and objective modifications to improve VLM performance on long captions and on compositional tasks. CLIP (Radford et al., 2021) offers strong general-purpose representations, but its short effective context window and weak compositional abilities expose clear limitations for complex captions (Zhang et al., 2024a; Kamath et al., 2023; Yamada et al., 2022). Recent efforts have introduced new training objectives (Yuksekgonul et al., 2022; Patel et al., 2024) and improved data quality (Doveh et al., 2023; Abbas et al., 2023) to boost performance.

2.1 Compositional Reasoning

Datasets Yuksekgonul et al. (2022) showed that many standard VLM benchmarks can be solved by detecting objects alone, without modeling their relationships. They thus proposed ARO, a benchmark for relational and attribute understanding, built on Visual Genome (Krishna et al., 2017), using GQA (Hudson and Manning, 2019) annotations. Complementary work explored compositional challenges through controlled linguistic or visual perturbations. VL-Checklist (Zhao et al., 2022), and CREPE (Ma et al., 2023) probe sensitivity to fine-grained structure such as object-attribute binding and spatial relationships, while Winoground (Thrush et al., 2022) evaluates sensitivity to word order using contrastive image-caption pairs with identical lexical content but different semantic interpretations. However, it is known to entangle compositional reasoning with world knowledge (Diwan et al., 2022).

A key limitation of earlier benchmarks is their reliance on implausible or unnatural negatives, which can allow models to exploit superficial cues (Hsieh et al., 2023). To address this, the authors introduced SugarCREPE, with more plausible negatives generated by large language models, and SugarCREPE++ (Dumpala et al., 2024) further extends this benchmark with paraphrased positives to better assess compositional generalization.

Models Models targeting compositionality typically modify the training objective to encourage sensitivity to relational structure. NegCLIP (Yuksekgonul et al., 2022) and CE-CLIP (Zhang et al., 2024b) introduce hard or structured negative captions during contrastive training. DAC (Doveh et al., 2023) and DCI (Urbanek et al., 2024) focus on improving caption quality through denser or more localized descriptions. Additional implementation details are provided in Appendix B.

2.2 Long-Caption Understanding

Standard CLIP-style models are trained on short, web-crawled captions and have a limited context

window of 77 tokens, often effectively attending to only the first 20–30 tokens (Zhang et al., 2024a). This motivates the use of long, dense captions both as a tool for training and for evaluating generalization under realistic linguistic complexity.

Datasets Several datasets provide long, dense captions that emphasize fine-grained descriptions, spatial relations, and contrastive distinctions. ImageInWords (Garg et al., 2024) and DOCCI (Onoe et al., 2024) consist of human-written captions targeting detailed scene understanding, with DOCCI explicitly designed to distinguish between visually similar images. Urban1k (Zhang et al., 2024a) and DCI/sDCI (Urbanek et al., 2024), derived from Visual Genome (Krishna et al., 2017), use synthetically generated or summarized captions to capture complex object relations while remaining compatible with CLIP-style models.

Larger-scale efforts include Localized Narratives (Pont-Tuset et al., 2020), which collects grounded human descriptions through synchronized narration and pointing, and synthetic approaches such as ShareGPT4V (Chen et al., 2024b) and LotLIP (Wu et al., 2024), which generate long captions at scale using large language models applied to standard CLIP pretraining corpora. These datasets vary widely in scale, grounding, and caption structure, reflecting different trade-offs between annotation quality and dataset size.

Models Long-caption models have primarily focused on extending the effective textual context available during training. LongCLIP (Zhang et al., 2024a) modifies the architecture to process longer input sequences, while DreamLIP (Zheng et al., 2024) emphasizes training on long, high-quality captions at scale without architectural changes. Detailed model descriptions in Appendix B.

3 Interplay of Compositionality and Long-Caption Understanding

We investigate the relationship between compositional reasoning and long-caption understanding in vision-language models (VLMs). Specifically, we ask: (Q1) Does training for compositionality improve a model's ability to interpret long, dense captions? and (Q2) Does training on long, structured captions promote compositional generalization? To answer these questions, we evaluate a set of off-the-shelf and custom-trained models using a unified benchmark suite, with all models tested in a zero-shot setting.

3.1 Experimental Setup

Compositionality Benchmarks We use Winoground and SugarCREPE++ (SC++) to evaluate compositionality. Although Winoground is an especially challenging benchmark, we include it as an upper bound for complex scene understanding, and, following Diwan et al. (2022), we report their proposed grouped scores in Appendix G to provide for fine-grained analysis. SC++, which encompasses and extends previous benchmarks, provides the most robust evaluation of compositional generalization. We also discuss the ARO benchmark in § 4.1, but exclude VL-CheckList because some of its images are no longer available, preventing full reproducibility[1].

Long-caption Retrieval Benchmarks To assess multimodal alignment under long, dense captions, we use zero-shot image-to-text and text-to-image retrieval as a proxy for understanding. We select datasets that contain detailed object-attribute bindings, spatial layouts, and nuanced descriptions beyond the scope of standard generic captions: Urban1K (Zhang et al., 2024a), sDCI (Urbanek et al., 2024), DOCCI (Onoe et al., 2024), and ImageInWords (Garg et al., 2024). Dataset statistics, presented in Table 8, highlight both the greater length of these captions and the increased complexity of their textual descriptions.

Main Models To study the interplay between compositional training and long-caption understanding, we evaluate a set of open-source VLMs. The compositional models we consider are NegCLIP, DAC_{LLM}, DCI_{P1}, and CE-CLIP, all trained with data and objectives designed to enhance compositional generalization (see Appendix B.) For long-caption understanding, we include LongCLIP-B and DreamLIP. As a baseline, we use the base CLIP model (ViT-B/32), which serves as the initialization for all other models in our study. We focus on CLIP-based models to enable controlled comparisons and isolate the effects of training data and objectives without architectural confounds.

Control Model Most models in our study fine-tune CLIP (ViT-B/32) using diverse datasets and training objectives. However, LongCLIP modifies

[1]This limitation has also been noted by Zhang et al. (2024b), who recommend avoiding VL-CheckList for new evaluations on their project website.

	Compositional Reasoning								Long-Caption Retrieval								
		SugarCrepe++							Urban1k		sDCI		DOCCI		IiW		
Model	WG	SA	RR	RO	RA	SO	Avg.		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	Avg.
CLIP	17.2	39.1	47.4	85.3	62.4	32.5	53.3		59.9	49.6	83.0	69.5	50.3	52.4	86.5	85.0	67.0
SigLIP	**18.6**	51.1	49.8	85.2	69.9	31.4	57.5		62.9	62.3	88.0	78.0	**70.3**	**70.8**	93.0	94.5	77.5
DAC_{LLM}	12.6	29.5	48.6	70.1	50.0	19.6	44.0		11.4	23.9	65.5	68.2	36.6	39.7	71.3	71.8	48.5
DCI_{P1}	12.1	41.8	39.5	80.7	56.6	38.0	51.3		29.7	43.0	71.4	70.8	42.9	46.3	71.3	75.3	56.3
CE-CLIP	12.3	41.6	52.6	85.7	68.3	33.3	56.3	→	53.5	65.0	82.0	74.8	43.1	55.4	73.3	85.3	68.1
NegCLIP	16.4	**57.5**	52.0	**92.1**	73.0	**43.9**	**63.7**		64.6	62.7	91.3	76.4	53.9	62.2	85.0	91.0	73.4
LongCLIP	14.7	40.8	48.4	89.1	65.6	29.6	54.7		77.9	77.7	86.4	74.6	61.4	69.3	90.8	**95.0**	79.1
DreamLIP	18.0	53.0	45.1	81.2	58.2	32.9	54.1	←	**79.8**	**79.6**	**94.7**	**82.0**	69.9	69.5	93.0	93.0	**82.7**
LSS	17.5	52.2	**53.4**	91.3	**74.9**	36.5	61.8		75.4	74.1	91.7	75.1	64.5	63.0	**94.0**	92.0	78.7

Table 1: **Compositional (left) and Long-Caption Retrieval (right) Performance Across Models.** Compositional reasoning is evaluated on Winoground (WG, full results in Appendix G) and SugarCrepe++ with subcategories: SA (Swap Attribute), RR (Replace Relation), RO (Replace Object), RA (Replace Attribute), and SO (Swap Object). Long-caption retrieval is assessed on Urban1K, sDCI, DOCCI, and IiW using image-to-text (I2T) and text-to-image (T2I) metrics. **NegCLIP** leads on compositional tasks, while **LongCLIP-B** excels at long-caption retrieval. Arrows indicate cross-capability generalization: → shows gains in long-caption understanding from compositional training; ← shows compositionality gains from long-caption training.

CLIP's architecture to extend the input context, and DreamLIP uses a larger backbone (see §2.2). These modifications may affect the interplay between long-caption training and compositionality in ways that are difficult to isolate. To control for these factors, we train the Long Story Short model (LSS), which finetunes CLIP (ViT-B/32) on the same ShareGPT4V data as LongCLIP, using standard contrastive loss and the original 77-token context window. This allows us to isolate the effect of long-caption data itself, independent of architectural changes, model size, or pretraining vs. finetuning. Importantly, LSS operates under the same architectural constraints as NegCLIP, DAC_{LLM}, and DCI_{P1}, enabling a direct and fair comparison between long-caption understanding and compositional training. We do not train LSS as a general-purpose model, but as a controlled intervention. Full details are in Appendix C.

3.2 Does Compositional Training Improve Long-Caption Understanding?

This question tests whether compositional properties evaluated by compositionality benchmarks are necessary for understanding long, dense captions with rich relational structure.

Table 1 reports compositional performance alongside downstream performance on long-caption retrieval for compositional models. We observe substantial variation among models on compositional benchmarks. DAC and DCI exhibit poor performance on both SC++ and WG, falling short even of the base CLIP model. Only NegCLIP consistently achieves a sizeable improvement over CLIP across all SC++ subcategories while maintaining a reasonable score on Winoground, indicating that contrastive training augmented with hard negatives can effectively induce compositionality.

Turning to long-caption retrieval, we observe a closely aligned pattern. NegCLIP, despite being trained only on short COCO captions, substantially outperforms CLIP across all long-caption benchmarks and closes much of the gap to LSS and LongCLIP. CE-CLIP exhibits intermediate transfer, while DAC and DCI fail to generalize, consistently underperforming even CLIP. Overall, long-caption retrieval performance mirrors compositional performance, with NegCLIP achieving the strongest results among the compositional models.

This parallel ordering across SC++ and long-caption retrieval provides strong evidence that **compositional reasoning generalizes to and supports long-caption understanding**. We further analyze the failure modes of DAC and DCI in Section 4.1. Having established the relationship between compositionality and long caption understanding in one direction, we turn to our second research question.

3.3 Do Long Captions Promote Compositionality?

The basis for this question is two-fold: (1) long, detailed captions provide a rich signal for learning fine-grained vision-language alignment, and (2) their richer compositional structure can foster

generalizable compositional abilities.

Table 1 presents the results. We first consider long-caption retrieval performance for LongCLIP, DreamLIP, and our control model, LSS. While performance varies between the retrieval datasets, all three models perform strongly. LSS, despite retaining CLIP's original 77-token context window, outperforms LongCLIP on several subtasks and falls short by only 0.4 points on average, suggesting limited benefit from LongCLIP's extended context. DreamLIP achieves the strongest overall retrieval performance, consistent with its larger backbone and full pretraining on long captions.

Next, we consider the performance of these models on compositionality benchmarks. LSS substantially improves over CLIP on SC++ and Winoground, nearly matching NegCLIP's average performance and surpassing it in some SC++ subcategories. LSS also achieves a high Winoground score, providing strong evidence that **training on long, grounded captions promotes compositional generalization**. DreamLIP performs well on WG and shows consistent, though smaller, gains over CLIP on SC++, while its larger model size and full pretraining introduce confounding factors.

In contrast, LongCLIP shows little improvement over CLIP on SC++ or Winoground, despite being trained on the same ShareGPT4V data as LSS. This divergence suggests that architectural constraints, specifically freezing early positional embeddings, can limit gains in compositional reasoning. We further analyze this effect in § 4.3. Figure 1 summarizes these results.

4 Limits of the Bidirectional Relationship

This section presents further analysis of the results discussed above, focusing on cases where the bidirectional relationship between compositionality and long-caption understanding does not hold uniformly. While these capabilities can reinforce each other, we observe substantial variation across models driven by data quality, training dynamics, and architectural constraints. We analyze these failure modes and examine how gains in compositional and long-caption understanding interact with other core vision-language capabilities.

4.1 The Case of DAC and DCI

The DAC_{LLM} and DCI_{P1} models were trained with compositional objectives and hard negatives on long-caption data (see Appendix B.), but they consistently underperform on the SC++ and Winoground benchmarks and on long-caption retrieval tasks (see Table 1.) This failure contrasts sharply with the success of NegCLIP and LSS, which each use one component from this training recipe and exhibit strong generalization.

	ARO Benchmark				
Model	VG-R	VG-A	COCO	Flickr	**SC++**
CLIP	59.8	63.0	47.3	58.5	53.3
SigLIP	34.8	55.9	32.7	40.7	57.5
DAC_{LLM}	81.3	73.9	**94.5**	**95.7**	44.0
DCI_{P1}	72.6	67.6	88.6	91.3	51.3
CE-CLIP	81.2	75.6	71.9	75.6	56.3
NegCLIP	**81.8**	72.1	82.5	86.7	**63.7**
LongCLIP	59.7	63.4	56.9	69.0	54.7
DreamLIP	51.2	**79.0**	52.0	49.8	54.1
LSS	62.0	65.7	37.7	46.0	61.8

Table 2: **Compositional Reasoning Performance on ARO and SugarCrepe++.** Compositional reasoning performance of various models on the ARO benchmark and the average SugarCrepe++ (SC++) score. Notably, models like DAC_{LLM} and DCI_{P1} almost saturate the ARO benchmark, yet their performance on SC++ often shows a poor correlation, indicating ARO's limitations in evaluating modern compositional abilities.

The original evaluations of DAC and DCI relied heavily on ARO, a legacy benchmark consisting of constrained, rule-based captions. As shown in Table 2, both models perform well on ARO, nearly saturating the benchmark. However, our analysis reveals a negative correlation between ARO and SC++ performance (Spearman $r = -0.37$), exposing a disconnect between ARO and more challenging contemporary evaluations. SC++ includes more natural distractors and paraphrased positives, requiring generalizable compositional reasoning. Strong ARO performance appears to reflect optimization for a narrow metric rather than true generalization. Still, this mismatch alone does not fully explain why DAC and DCI perform so poorly on the tasks they were trained to address.

We posit that the failure to generalize stems from two core issues. First, the training captions used by DAC and DCI may be of low quality. DAC and DCI rely heavily on synthetic captions, either blind LLM-generated expansions or disjoint region-level descriptions, that may often lack fluency, cohesion, or visual grounding. These deficiencies limit the ability of a model to learn grounded, compositional alignments—even in the presence of hard negatives and contrastive loss. Second, both DAC and DCI

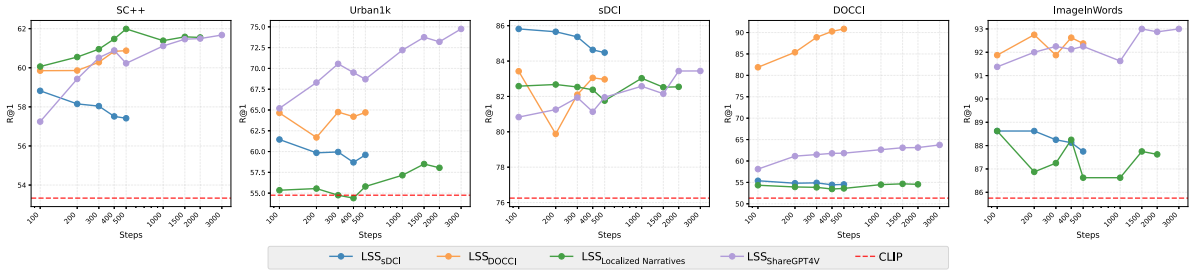

Figure 2: **Training Dynamics Across Long-Caption Datasets**. Evolution of performance for models trained on four long-caption datasets. Results are reported for both long-caption retrieval (Urban1K, DOCCI, sDCI, IiW) and compositional benchmark, SC++. Models trained on ShareGPT4V and DOCCI show consistently stronger generalization, highlighting the role of caption quality and grounding beyond dataset scale.

apply LoRA-based updates to only a limited subset of model parameters. This lightweight adaptation restricts the models' capacity to internalize compositional structure beyond surface-level patterns. In contrast, NegCLIP and LSS were fully finetuned, allowing deeper representational shifts.

Together, these findings highlight the importance of not just what models are trained on, but also how they are trained. Compositional objectives and long captions alone are not sufficient; sufficient parameter adaptation and data quality are critical to realizing the intended generalization. Next, we examine the role of the training data further.

4.2 Are All Long-Caption Datasets the Same?

To study how different types of long-caption data affect compositional reasoning and long-caption retrieval, we train additional LSS models on the sDCI Train, DOCCI Train, Localized Narratives datasets, using the same experimental setup from §3.1). These datasets vary in scale and linguistic complexity, range from 7K to over 1M examples in size, and span both synthetic and human-written captions. Table 3 shows detailed dataset statistics. This controlled intervention allows us to systematically examine how properties like grounding, compositional complexity, and size influence generalization.

We show the results in Figure 2 (see Appendix H). all LSS models outperform the CLIP baseline on long-caption retrieval and exhibit some degree of compositional transfer. However, performance varies notably: models trained on ShareGPT4V and DOCCI consistently outperform those trained on sDCI and Localized Narratives, both in the course of training and at convergence. This indicates that generalization depends not just on scale, but also on the quality and structure of the training data. We now discuss each training dataset in turn.

ShareGPT4V. LSS$_{\text{ShareGPT4V}}$ achieves the strongest results overall, excelling in both long-caption retrieval and compositional benchmarks like SC++. This success stems from a combination of favorable dataset properties (Table 8) rather than any single factor. At 1.2M examples, ShareGPT4V is the largest scale dataset, but more critically, it achieves near-complete vocabulary coverage (87.72% of CLIP's tokenizer vocabulary), far exceeding the ∼25% coverage of the other datasets. The average caption lengths are far-beyond the 77-token limit of CLIP, ensuring the model trains the full range of positional embeddings. Interestingly, ShareGPT4V exhibits only moderate syntactic complexity (Yngve depth: 45.70), which is significantly lower than the other datasets. This suggests that the combination of scale, comprehensive vocabulary coverage, and appropriate caption length compensates for moderate structural complexity, enabling robust generalization on both compositional and long-caption benchmarks.

DOCCI. Despite the small size of DOCCI (14.6K captions), LSS$_{\text{DOCCI}}$ achieves strong performance comparable to LSS$_{\text{ShareGPT4V}}$ across long-caption retrieval tasks. As shown in Table 8, its strength lies in being human-written, visually grounded, and contrastive by design, explicitly crafted to distinguish between similar images. With high syntactic complexity (Yngve: 74.55) and long captions (122 tokens), DOCCI demonstrates that careful curation and purposeful annotation strategies can rival large-scale data collection, achieving comparable vocabulary coverage (26.96%) despite its size.

sDCI. LSS$_{\text{sDCI}}$ achieves strong performance on the sDCI test set, but performs less consistently on

other benchmarks. Although the sDCI train set shows the highest syntactic complexity (Yngve: 94.07), this may stem from LLM-generated caption summarized from multiple captions exceeding 1000 words, which may artificially inflate syntactic complexity without good semantic coherence or visual grounding. Its small image set (7.6K unique images) may also lead to overfitting, as indicated by the declining performance on SC++ and Urban1K with more training steps. This highlights that complexity alone is not enough without grounding and sufficient visual diversity.

Localized Narratives. With 489K captions, LN is the only large-scale human-annotated dataset available. LSS_{LN} shows the fastest early gains on SC++ among the long-caption datasets (Figure 2), likely thanks to a moderate-high Yngve complexity (61.70). However, the model lags behind on all the long-caption retrieval benchmarks. This can be attributed to caption length: LN captions average only 30 words (Table 8), well below CLIP's 77-token limit and far shorter than the other DOCCI, sDCI and ShareGPT4V. Moreover, LN has the lowest vocabulary coverage (24.34%). LN provides enough signal to improve compositionality, but the captions are too short on average to support strong long-caption understanding.

Our analysis shows that dataset effectiveness depends on the interaction of multiple properties—scale, vocabulary coverage, caption length, syntactic complexity, and annotation quality—rather than any single factor. High-performing datasets like ShareGPT4V and DOCCI succeed through complementary strengths: large scale and broad coverage in the former, careful grounding and contrastive structure in the latter. In contrast, sDCI and Localized Narratives highlight key pitfalls: overly complex or synthetic captions without grounding, and human-annotated but short captions with limited vocabulary. Ultimately, strong generalization emerges not from any one property, but from a balanced design that aligns linguistic richness with visual grounding and sufficient sequence length.

4.3 The Case of LongCLIP

Our main results in Table 1 revealed a surprising finding: despite having a very different context window size, $\text{LSS}_{ShareGPT4V}$ and LongCLIP, trained on the same ShareGPT4V dataset, show comparable performance on long-caption retrieval, and the former even outperforms the latter on compositionality tasks. At first sight, this is counterintuitive, since LongCLIP is expected to benefit from a major architectural advantage, which allows the model to process up to 248 tokens (compared to 77 for LSS.) However, the differences are not just in architecture but also in training procedure. LSS operates within CLIP's original 77-token limit, updating all parameters during training. LongCLIP, by contrast, uses an extended context window but freezes the first 20 positional embeddings and applies reduced updates to positions 20 to 77. These modifications aim to preserve the strong generalization capabilities of the base CLIP model. Yet, they may also interfere with the learning of new skills in this part of the context window.

Dataset	Images	Captions	Avg. Length	Vocab. Covered (%)
$\text{sDCI}_{\text{Train}}$	7.6×10^3	8.3×10^4	40 ± 12	29.29
$\text{DOCCI}_{\text{Train}}$	1.5×10^4	1.5×10^4	122 ± 45	26.96
LN	4.9×10^5	4.9×10^5	30 ± 17	24.34
ShareGPT4V	1.2×10^6	1.2×10^6	144 ± 39	87.72

Table 3: **Training dataset statistics.** Dataset size, average caption length, and vocabulary coverage with respect to CLIP's tokenizer. ShareGPT4V stands out for its scale, long captions, and near-complete vocabulary coverage. Additional dataset statistics in Table 8 (including syntactic complexity, word ranges, source types, and evaluation datasets).

To isolate this effect, and eliminate the confounding factor of longer input length, we evaluate a truncated version of LongCLIP: LongCLIP_{70}, which limits inputs to 70 words (approximately $\sim$77 tokens). As shown in Figure 3, performance on long-caption retrieval for LongCLIP_{70} drops sharply, and $\text{LSS}_{\text{ShareGPT4V}}$ now dominates on both compositionality and long-caption understanding. The main limitation of LongCLIP thus appears to stem from its training constraints. Freezing the first 20 positional embeddings on the assumption that they are already well-optimized for vision–language alignment, limits the ability to learn new relational patterns. As a result, it preserves a bag-of-words–like processing, particularly in the early input positions where most compositional benchmarks operate. In the next section, we explore this trade-off in more depth.

4.4 Trade-offs for Better Compositionality and Long-Caption Understanding

To better understand how compositionality and long-caption training affect general alignment, we

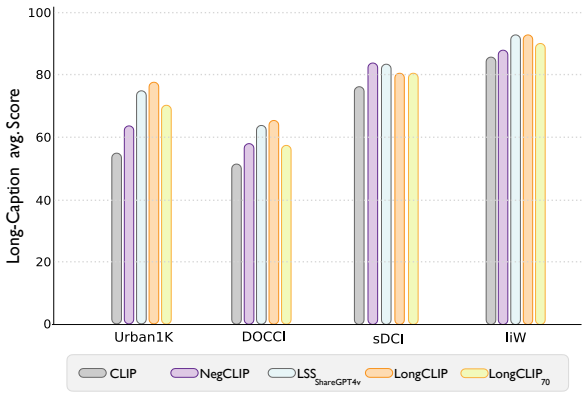

Figure 3: **Long-caption Retrieval Performance**. LSS achieves strong performance on all benchmarks, with LongCLIP outperforming by a small margin despite its three times larger input token processing capacity. When truncating LongCLIP to 70 words, LSS closes the gap and outperforms it across all benchmarks. These results demonstrate that full parameter adaptation enables more effective long-caption understanding.

	Zero-Shot Classification			Retrieval			
				COCO		Flickr30k	
Model	**C10**	**C100**	**IN1K**	I2T	T2I	I2T	T2I
CLIP	89.8	65.1	63.1	50.4	30.2	78.6	59.0
SigLIP	92.4	**72.3**	**76.0**	**65.7**	**47.8**	**88.9**	74.7
DAC_{LLM}	90.4	64.1	51.1	33.7	37.7	53.1	64.9
DCI_{P1}	87.1	58.0	53.3	20.5	21.4	55.9	44.0
CE-CLIP	85.9	60.2	50.0	55.3	46.9	74.9	68.3
NegCLIP	88.9	63.2	61.0	59.3	44.8	85.1	70.9
LongCLIP	91.3	69.5	66.9	57.2	40.6	86.2	70.7
DreamLIP	**92.7**	67.0	55.7	57.6	47.7	84.7	**75.0**
LSS	88.9	65.8	60.8	57.2	38.9	83.0	68.4

Table 4: **Zero-shot Performance on General Vision-Language Tasks.** We report classification accuracy on CIFAR10 (C10), CIFAR100 (C100), and ImageNet-1K (IN1K), and Recall@1 for image-text retrieval on COCO and Flickr. LongCLIP performs strongly across all tasks, indicating that freezing initial positional embeddings preserves general alignment and even enhances it through the remaining positions. While NegCLIP and LSS underperform on classification, both achieve strong retrieval scores.

evaluate all models on standard vision-language benchmarks. We assess classification on CIFAR10, CIFAR100 and ImageNet using prompt templates from prior work (Radford et al., 2021), and retrieval (measured in Recall@1) on the COCO and Flickr30k test sets. Results are shown in Table 4.

LongCLIP achieves the strongest overall performance, particularly on classification and short-caption retrieval. This attests to the effectiveness of preserving CLIP's alignment via the freezing of positional embeddings. Since many COCO and Flickr30k captions exceed CLIP's functional 20-token window (Table 8), LongCLIP gains a natural retrieval advantage. NegCLIP achieves the highest recall on most short-caption retrieval tasks, demonstrating that compositional reasoning can enhance general vision-language understanding. However, this advantage may partly reflect in-distribution bias, as NegCLIP is trained on COCO, which appears in the evaluation set. Despite strong retrieval performance, NegCLIP shows degraded classification accuracy compared to base CLIP. We observe the same pattern in our $LSS_{ShareGPT4V}$ model: it consistently improves over base CLIP on all retrieval tasks but underperforms on classification, particularly on ImageNet. A likely explanation is the mismatch between training and evaluation distributions—these models are trained on longer, relational captions, but evaluated using fixed, template-based prompts (e.g., "a photo of a dog"), which favor alignment patterns seen during CLIP's original training.

These results point to a nuanced trade-off in VLM design. Constraints like frozen positional embeddings, like those used in LongCLIP, can preserve general performance, and even support improved performance when combined with architectural modifications or high-quality data. Compositional and long-caption understanding, as seen in LSS and NegCLIP, boosts retrieval but tends to degrade classification accuracy. This indicates that while compositionality and long-caption understanding benefit each other, both properties may come at the cost of general-purpose alignment, highlighting the importance of aligning training strategies with intended use cases.

5 Conclusion

We investigated the relationship between compositional reasoning and long-caption understanding in contrastive vision-language models. Our results reveal a bidirectional link: compositional training improves performance on long-caption retrieval, while training on long, complex captions fosters compositional generalization. However, we find that neither compositional objectives nor long captions are sufficient on their own. Robust generalization emerges only when combined with sufficient parameter adaptation and high-quality training data consisting of grounded, comprehensive captions with syntactic and lexical diversity.

Compositional generalization, whether achieved via targetted objectives or training data, is key to developing models that generalize across caption complexity. As models are increasingly applied to open-ended, real-world tasks, supporting robust generalization across diverse and structured language should be a central goal—one that requires choosing the right benchmarks, model design and training strategies.

Limitations

Our work highlights the benefits of long-caption training but focuses primarily on general performance, compositionality, and long-caption retrieval on contrastive visual-language models. We do not evaluate generative VLMs, which involve additional factors such as autoregressive decoding, alignment procedures, and cross-attention dynamics. Complex caption phenomena such as temporal or causal reasoning are not addressed. We also rely on long-caption retrieval as a proxy for understanding; more targeted benchmarks are needed for deeper analysis, potentially involving generative evaluation or fine-grained VLM probing. Finally, we do not explore combinations of training losses or architectural modifications; nonetheless, our results indicate that substantial gains are possible even within the original CLIP input constraints. While we observe strong correlations between compositional training and long-caption understanding, our analysis does not fully isolate the underlying causal mechanisms. Future work should investigate the specific roles of vocabulary coverage, syntactic complexity, and visual grounding through targeted mechanistic interventions.

Acknowledgments

This work was supported by research grant (VIL53122) from Villum Fonden. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC Development Access call (ID:EUHPC_D12_071).

References

Amro Kamal Mohamed Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*.

Dzmitry Bahdanau, Harm de Vries, Timothy J O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. 2019. Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. 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.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, and 1 others. 2023. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Haoxin Li and Boyang Li. 2025. Enhancing vision-language compositional understanding with multimodal synthetic data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24849–24861.

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

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

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, and 1 others. 2024. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pages 291–309. Springer.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and 1 others. 2024. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, and 1 others. 2023. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658.

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

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*, pages 647–664. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and 1 others. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Dídac Surís, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. 2020. Learning to learn words from visual scenes. In *European Conference on Computer Vision*, pages 434–452. Springer.

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.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709.

Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. 2024. Lotlip: Improving language-image pre-training for long text understanding. *arXiv preprint arXiv:2410.05249*.

Yutaro Yamada, Yingtian Tang, Yoyo Zhang, and Ilker Yildirim. 2022. When are lemons purple? the concept association bias of vision-language models. *arXiv preprint arXiv:2212.12043*.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

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

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

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

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. 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*, pages 30–37.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. 2024. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pages 73–90. Springer.

A General-Purpose VLM Evaluation

Pretraining VLMs using image-text pairs and contrastive loss functions (Oord et al., 2018) has proven highly effective for learning aligned multimodal representations (Radford et al., 2021; Jia et al., 2021; Pham et al., 2023), demonstrating strong generalization in zero-shot settings, where no task-specific finetuning is applied. Pretrained models are commonly evaluated on zero-shot classification datasets like CIFAR (Krizhevsky et al., 2009), as a benchmark for coarse-grained object recognition, and ImageNet-1k (Russakovsky et al., 2015), evaluating fine-grained classification across 1,000 categories. Oxford-IIIT Pets (Parkhi et al., 2012), Stanford Cars (Krause et al., 2013), and Food101 (Bossard et al., 2014) focus on fewer but more specific classes. Cross-modal retrieval benchmarks, such as MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015), contain short and general captions, evaluating alignment between visual and textual modalities through image-to-text and text-to-image matching. However, these benchmarks primarily test coarse-level alignment or object recognition, offering limited insight into compositional understanding.

B Compositional and Long-Caption Models

Compositional Models Various data augmentations and objective modifications have been proposed to improve the compositionality of VLMs. NegCLIP (Yuksekgonul et al., 2022) extends the contrastive loss with hard negative captions, encouraging finer discriminative alignment. Hard negatives are mined through nearest-neighbour search

and synthetically created through word order perturbations. Densely Aligned Captions (Doveh et al., 2023, DAC) targeted both data quality and training objectives. Based on the intuition that more descriptive captions can enhance compositionality, they introduced two methods for rewriting CC3M captions into more descriptive and semantically rich versions: 1) using an LLM to expand short captions with plausible scene-level details, and 2) using SAM (Kirillov et al., 2023) to segment objects and generate short captions for each region. Two models were trained, DAC_{LLM} and DAC_{SAM}, using LoRA(Hu et al., 2022), with DAC_{LLM} performing better on ARO. Urbanek et al. (2024) introduced a human-annotated dataset in which annotators were asked to describe automatically segmented regions of an image in detail. These region-level captions were then summarized to fit within the context window of CLIP, and used to fine-tune a vision-language model. They trained DCI_{P1} with both contrastive and negative loss showing strong performance on ARO and VL-Checklist. CE-CLIP (Zhang et al., 2024b) expands the hard-negative approach for compositionality by generating multiple targeted negatives for each image, focusing on relationships, attributes, and actions. The model is trained with the standard contrastive loss plus two additional objectives: (1) an intra-modal contrastive loss comparing correct captions to hard negatives, and (2) a cross-modal contrastive loss enforcing higher similarity between image–correct pairs than image–negative pairs. SPARCL (Li and Li, 2025), in contrast, introduces both positive and negative captions and images using a text-to-image generative pipeline.[2]

Long-Caption Models LongCLIP (Zhang et al., 2024a) focuses on long-caption processing by extending the textual context window from 77 to 248 inputs by interpolating positional embeddings. To retain the strengths of the original CLIP, the first 20 positional embeddings are frozen, while the remaining base encoder up to the token 77 are interpolated with a heavier weighting toward the pretrained positions. The rest of the model is fully trained from a CLIP initialization. DreamLIP (Zheng et al., 2024) explores the impact of training with long, high-quality captions. The authors recaption 30M images and train using a global multi-positive contrastive loss, pairing each image with multiple subcaptions, and a fine-grained loss aligning image subpatches with their corresponding subcaptions. Unlike other models discussed in this paper, DreamLIP uses a larger CLIP backbone (ViT-B/16 rather than ViT-B/32) and conducts a data-scaling study, training multiple model variants with increasing data sizes.

[2]Weights were not available at the time of writing.

Model	WarmUp	LR	Steps	Epochs	Checkpoints
sDCI	5	5e-6	500	70	[100, 200, 300, 400, 500]
DOCCI	15	5e-6	500	35	*PREV*
LN	50	3e-6	2000	4	[*PREV*, 1000, 1500, 2000]
ShareGPT4V	150	3e-6	3000	2.5	[*PREV*, 3000]

Table 5: **Training Parameters.** *PREV* includes all checkpoints from the model on the previous row, and *LR* denotes the learning rate. We report the number of training steps and their approximate equivalent in epochs.

C Training Parameters

In this section, we present the training parameters for our models. Models are named after the datasets on which they were trained.

We fine-tune CLIP models on each dataset, fixing the batch size to 1024 for all runs and adjusting the number of training steps accordingly. Training is performed using 4× A100 GPUs to accommodate the ViT-B/32 version of CLIP from HuggingFace.[3] Visual and textual input processing follows the default parameters of the pretrained CLIP models, as specified in the HuggingFace model card. The longest training run required 8 hours. Detailed training parameters are shown in Table 5.

D Dataset Statistics

Table 8 presents an overview of the datasets used in this study, in terms of the caption lengths, vocabulary coverage, and textual complexity, as estimated by Yngve metric.

E Pretraining vs. Fine-tuning

DreamLIP provides an opportunity to compare the effects of full pretraining with long captions against finetuning on long-caption data. Although the setups are not perfectly aligned, we fine-tune a larger CLIP model (ViT-B/16) using the same training recipe as $\text{LSS}_{\text{ShareGPT4V}}$. The results are summarized in Table 6.

We observe comparable performance between DreamLIP and our finetuned model on retrieval tasks. However, they differ notably on other

[3]Pretrained models will be released under an open-source license.

benchmarks. DreamLIP achieves the strongest Winoground score overall, whereas finetuning on a larger backbone does not significantly boost Winoground performance. In contrast, our fine-tuned LSS/16 model outperforms DreamLIP on SC++. The bigger differences emerge in image classification. Our fine-tuned model achieves the highest classification accuracy among all evaluated models, whereas DreamLIP underperforms even CLIP-B/32 despite its larger backbone. This may reflect the fact that pretraining exclusively on long captions introduces a distribution shift that harms zero-shot classification performance. For short retrieval, both models perform similarly.

F SigLIP Results

To complement our main results and validate that our findings generalize beyond CLIP-based architectures, we report results for SigLIP (Zhai et al., 2023), a vision–language model trained with a sigmoid loss on the large-scale WebLI dataset (Chen et al., 2022). As shown in Table 6, SigLIP demonstrates stronger generalization on classification and short retrieval tasks compared to CLIP-based models, DreamLIP and LSS (ViT-B/16).

Despite the absence of explicit long-caption supervision, SigLIP achieves performance comparable to DreamLIP and LSS/16 on long-caption retrieval benchmarks and performs strongly on Winoground and SC++. This consistency across tasks supports our central conclusion that long-caption understanding and compositional reasoning are mutually reinforcing capabilities. Importantly, SigLIP's results, achieved with a different objective function, training data, and architecture, demonstrate that these relationships are not artifacts of CLIP's design but rather may reflect general principles of vision–language learning.

G Full Winoground Results

We report the complete Winoground results, including separate scores for Text, Image, and Group. In addition, we group results by the tags proposed by Diwan et al. (2022), which break the benchmark into finer-grained subcategories and allow for deeper analysis (see Table 7).

Across the full dataset (Overall columns), LSS achieves the highest Text score and competitive performance on Image and Group. When grouped by tags, LSS consistently shows strong Text performance, often ranking first or close to the best-performing model. Interestingly, vanilla CLIP achieves strong results on several individual groups, highlighting that Winoground remains challenging for current representational learning approaches. These results reinforce that, while progress is being made, compositional reasoning in VLMs is still far from solved.

H Compositionality and Long-Caption Retrieval Results for LLS Models.

We provide a detailed breakdown of the performance for our LSS model variants. Table 9 presents the complete results for LSS models trained on different datasets of image and long-caption pairs. The reported metrics cover both long-caption retrieval and compositionality benchmarks, offering a comprehensive overview that allows for a direct comparison of how different training data affects final model capabilities.

	Winoground			Long Retrieval			
Model	Group	Image	Text	Urban	sDCI	DOCCI	IiW
Siglip	10.3	12.8	32.8	62.6	83.2	70.6	93.8
DreamLIP	10.8	15.0	28.3	79.7	88.3	69.7	93.0
LSS (ViT-B/16)	8.8	11.3	30.8	81.3	83.7	67.8	94.8
	Classification			**Short Retrieval**			
	CIFAR10	CIFAR100	ImageNet	COCO	Flickr30k	SC++	
Siglip	92.4	72.3	76.0	56.8	81.8	57.5	
DreamLIP	92.7	67.0	55.7	52.7	79.9	54.1	
LSS (ViT-B/16)	91.4	68.0	65.1	51.8	81.2	60.4	

Table 6: **Generalization Across Architectures.** DreamLIP (pretrained on long captions) and LSS/16 (finetuned on long captions) achieve comparable performance across benchmarks. SigLIP, despite lacking explicit long-caption training, matches these models on long-caption retrieval and performs strongly on compositional benchmarks. These results reinforce that the observed relationship between long-caption understanding and compositionality generalizes across architectures and training paradigms.

Model	Overall			Non Compositional			Unusual Image			Visually Difficult			Unusual Text			Ambiguously Correct			Complex Reasoning			NoTag		
	T.	I.	G.	T.	I.	G.	T.	I.	G.	T.	I.	G.	T.	I.	G.	T.	I.	G.	T.	I.	G.	T.	I.	G.
CLIP	31.3	11.3	9.0	**76.7**	40.0	36.7	**25.0**	8.9	5.4	15.8	0.0	0.0	**36.0**	14.0	10.0	30.4	**15.2**	**15.2**	24.4	6.4	**3.8**	32.0	11.6	9.3
SigLIP	32.8	12.8	10.3																					
DAC_{LLM}	22.5	10.3	4.8	50.0	26.7	20.0	16.1	10.7	3.6	10.5	10.5	7.9	20.0	10.0	2.0	19.6	10.9	4.3	19.2	**10.3**	**3.8**	23.8	9.9	3.5
DCI_{P1}	20.8	10.3	5.3	53.3	26.7	23.3	23.2	5.4	5.4	21.1	2.6	2.6	18.0	4.0	2.0	15.2	10.9	4.3	19.2	**10.3**	**3.8**	18.0	11.0	4.7
CE-CLIP	19.5	12.0	5.3	36.7	33.3	13.3	12.5	**16.1**	**8.9**	5.3	7.9	0.0	18.0	16.0	6.0	23.9	10.9	6.5	17.9	7.7	2.6	20.3	11.0	4.7
NegCLIP	30.3	11.0	8.0	66.7	30.0	26.7	17.9	8.9	3.6	10.5	2.6	2.6	**36.0**	10.0	8.0	28.3	4.3	4.3	21.8	9.0	5.1	35.5	12.2	8.7
LongCLIP-B	28.5	8.8	7.3	66.7	40.0	**40.0**	**25.0**	5.4	5.4	**28.9**	2.6	2.6	28.0	14.0	12.0	32.6	8.7	8.7	25.6	0.0	0.0	26.7	8.7	5.8
DreamLIP	28.2	**15.0**	**10.8**	46.7	**43.3**	33.3	23.2	10.7	7.1	23.7	**13.2**	**10.5**	24.0	**18.0**	**14.0**	**34.8**	13.0	8.7	**28.2**	9.0	5.1	26.7	**17.4**	**13.4**
LSS	**33.3**	11.8	7.5	66.7	26.7	26.7	**25.0**	8.9	3.6	18.4	0.0	0.0	32.0	8.0	8.0	32.6	**15.2**	13.0	**28.2**	7.7	2.6	**36.0**	14.0	8.1

Table 7: **Full Winoground Performance and Tag Breakdown.** The best-performing model on grouped results varies across tags. LSS demonstrates consistently strong textual understanding, achieving the highest overall score and leading in several categories, including complex reasoning. These results suggest that training with long captions helps improve performance on this benchmark.

Dataset	Images	Captions	Avg. Length	Word Range	Vocab. Covered (%)	Yngve	Source
$sDCI_{Train}$	7599	82 785	40 ± 12	4 - 70	29.29	94.07	Mixed
$DOCCI_{Train}$	14 647	14 647	122 ± 45	28 - 518	26.96	74.55	Human
LN	489 000	489 000	30 ± 17	1 - 226	24.34	61.70	Human
ShareGPT4v	1 200 000	1 200 000	144 ± 39	19 - 507	87.72	45.70	Synthetic
$COCO_{Test}$	5000	24 855	10 ± 3	6 - 43	14.75	24.53	Human
$Flickr30k_{Test}$	1000	4999	12 ± 5	2 - 68	8.99	34.64	Human
Urban1k	1000	1000	107 ± 10	74 - 179	10.75	69.27	Synthetic
IiW_{Test}	400	400	217 ± 83	45 - 480	11.53	119.22	Mixed
$sDCI_{Test}$	206	2236	41 ± 12	11 - 67	7.67	95.64	Mixed
$DOCCI_{Test}$	5200	5200	123 ± 46	28 - 518	19.29	73.95	Human

Table 8: **Dataset Statistics for Training and Evaluation**. Datasets above the dashed line are used for training, while those below are used for testing. The 'Word Range' column indicates the words in the shortest and longest caption in the dataset and the 'Vocabulary Covered (%)' column indicates the percentage of tokens found in the respective datasets that are also present in the tokenizer training vocabulary of CLIP. The 'Yngve' metric quantifies syntactic complexity, with higher values indicating deeper left-branching structures (Yngve, 1960). Many datasets exceed CLIP's 77-token input limit, especially those designed for long-caption understanding.

Model	Compositional Reasoning								Long-Caption Retrieval								
		SugarCrepe++						Urban1k		sDCI		DOCCI		IiW			
	WG	SA	RR	RO	RA	SO	Avg.	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	Avg.
LSS_{sDCI}	**18.5**	47.0	51.0	90.5	69.4	29.2	57.4	61.2	58.0	90.3	**78.7**	54.7	54.3	87.5	88.0	71.6
LSS_{DOCCI}	18.0	51.2	**56.3**	90.1	74.4	32.4	60.9	67.2	62.2	89.3	76.6	**90.6**	**91.1**	93.3	91.5	82.7
LSS_{LN}	12.9	53.5	52.9	**93.1**	73.6	34.7	61.6	57.8	58.3	88.8	76.3	54.1	55.0	88.3	87.0	70.7
$LSS_{ShareGPT4V}$	17.5	52.2	53.4	91.3	**74.9**	**36.5**	**61.7**	75.4	74.1	**91.7**	75.1	64.5	63.0	**94.0**	92.0	78.7

Table 9: **Compositional (left) and Long-Caption Retrieval (right) Performance Across Models.**