Auto-Comp: Scalable Controlled Synthetic Benchmarks for VL Compositionality

Cristian Sbrolli
Politecnico di Milano
cristian.sbrolli@polimi.it

Toshihiko Yamasaki
The University of Tokyo
yamasaki@cvm.t.u-tokyo.ac.jp

Matteo Matteucci
Politecnico di Milano
matteo.matteucci@polimi.it

Abstract

Modern vision-language models struggle with basic compositional reasoning, often confusing "a red cube and a blue sphere" with "a blue cube and a red sphere". Disentangling the visual and linguistic roots of these failures is a fundamental challenge for robust evaluation. To enable fine-grained, controllable analysis, we introduce Auto-Comp, a fully automated and synthetic pipeline for generating scalable benchmarks. Its controllable nature is key to dissecting and isolating different reasoning skills. Auto-Comp generates paired images from Minimal (e.g., "a monitor to the left of a bicycle on a white background") and LLM-generated Contextual captions (e.g., "In a brightly lit photography studio, a monitor is positioned to the left of a bicycle"), allowing a controlled A/B test to disentangle core binding ability from visio-linguistic complexity. Our evaluation of 20 VLMs on novel benchmarks for color binding and spatial relations reveals universal compositional failures in both CLIP and SigLIP model families. Crucially, our novel "Confusion Benchmark" reveals a deeper flaw beyond simple attribute swaps: models are highly susceptible to low-entropy distractors (e.g., repeated objects or colors), demonstrating their compositional failures extend beyond known bag-of-words limitations. we uncover a surprising trade-off: visio-linguistic context, which provides global scene cues, aids spatial reasoning but simultaneously hinders local attribute binding by introducing visual clutter. We release the Auto-Comp pipeline to facilitate future benchmark creation, alongside all our generated benchmarks.

1. Introduction

Contrastive Vision-Language Models (VLMs) have demonstrated remarkable capabilities in connecting text and images, enabling applications ranging from detailed image captioning to segmentation and image generation [9–11]. By learning from vast, web-scale datasets, these models build powerful, unified embedding spaces for images and text. However, a critical failure point persists in their ability to perform robust compositional reasoning, the capacity to understand and correctly bind elements within a scene based on their attributes and relationships. For instance, a state-of-the-art VLM may correctly identify the presence of a "red cube" and a "blue sphere" in an image, yet fail to distinguish it from an image containing a "blue cube" and a "red sphere". This failure in compositional binding is not a minor flaw; it reveals a fundamental weakness in achieving true, human-like scene understanding and limits the reliability of these models in high-stakes applications.

Evaluating this compositional capability is notoriously difficult. Benchmarks built on real-world images [2, 4, 6, 7], while ecologically valid, are inherently noisy, making it impossible to isolate specific reasoning failures from confounding visual variables. Conversely, existing synthetic benchmarks that offer more control often lack the photorealism needed to fairly evaluate modern VLMs or use simplistic, template-based language that fails to mirror real-world linguistic diversity [14]. This leaves a critical question unanswered: do these models have a core compositional ability that is simply brittle to the *visio-linguistic complexity* of realistic scenes, or is their understanding fundamentally flawed? Consequently, a significant gap exists for a diagnostic tool that can systematically disentangle these factors.

To solve this disentanglement challenge, we introduce **Auto-Comp**, a fully automated pipeline for generating scalable, photorealistic benchmarks. The core innovation of our framework is a parallel generation process designed specifically for controlled A/B testing. For each compositional concept, Auto-Comp creates two distinct visio-linguistic conditions: (1) a *Minimal* condition, featuring simple, template-based captions and objects isolated on a white background to test a model's core binding ability; and (2) a *Contextual* condition, where an LLM rewrites the caption into natural language and a text-to-image model renders the objects within a realistic scene. This parallel structure allows us to precisely measure how performance changes when moving from a sterile to a complex environment. To ensure data quality, the entire pipeline incorporates a rigorous two-stage validation process and is designed to be general, enabling the generation of benchmarks for any user-defined compositional skill. We make the following contributions:

Table 1. Comparison of compositional VLM benchmarks. Our work, Auto-Comp, is the first to combine photorealistic images with a fully automatic, concept-driven generation pipeline.

Benchmark	# of Samples	Realistic Images	Realistic Text	Concept-driven Generation	Fully Automatic
CREPE [7]	370,000+	✓	✗	✗	✓
SugarCrepe [4]	7,512	✓	✓	✗	✗
SugarCrepe++ [4]	4,757	✓	✓	✗	✗
Winoground [17]	400	✓	✓	✗	✗
CIVET [14]	∼35,000	✗	✓	✓	✓
Auto-Comp-CP (Ours)	**175,221**	✓	✓	✓	✓

- We present **Auto-Comp**, a fully automated and scalable pipeline for generating and validating high-quality, photorealistic synthetic datasets, enabling the creation of custom evaluation data for any user-defined compositional skill.
- We propose a novel A/B evaluation framework using parallel *Minimal* and *Contextual* that allows to isolate core visual binding failures from visio-linguistic contextual complexity. We discover a common vulnerability to low-entropy distractors that goes beyond the known "bag-of-words" issues identified by previous studies.
- Using Auto-Comp, we generate and release Auto-Comp-CP, comprising two comprehensive benchmarks for *Color* (N=1, 2, 3) and *Position* (N=2, 3) binding. we introduce two novel hard-negative evaluation schemes: the *Swap Benchmark* and the more challenging *Confusion Benchmark*.
- We conduct an extensive analysis of over 20 VLMs, revealing a clear performance hierarchy where SigLIP models [18, 20] outperform CLIP [9] and uncovering a surprising trade-off: visio-linguistic context aids spatial reasoning but hinders attribute binding.

Our work provides an automatic, powerful and scalable pipeline for the community, enabling a deeper understanding of current VLM limitations, paving the way for the development of more robust models, and providing a flexible tool for future investigations into compositional skills.

2. Related Work

Our research is situated at the intersection of VLM evaluation, compositional reasoning, and synthetic data generation. We position our contributions in relation to the key paradigms of prior work, summarized in Table 1.

Data-driven Benchmarks on Real Images. A primary approach to compositional VLM evaluation uses pre-existing datasets, where innovation focuses on curating the negative texts. The large-scale CREPE benchmark [7], for instance, starts from real image-caption pairs and programmatically swaps attributes to create hard negatives. While automated, this process can produce linguistically awkward captions, creating an artifact that models may exploit. Addressing this, SugarCrepe [4] and its successor, SugarCrepe++ [2], leverage LLMs to generate fluent, realistic negative captions, significantly improving text quality. However, despite achieving realism in both modalities, these benchmarks required human evaluation and remain fundamentally data-driven, not concept-driven. They are constrained by the content of pre-existing images and thus cannot control and isolate objects and distracting elements.

Concept-driven Benchmarks with Synthetic Visuals. To gain composition control, a second line of research uses synthetic visuals. The CIVET [14] framework is a prime example, programmatically placing real-object cutouts onto a simple grid to create scenes. While this methodology offers precise, concept-level control, it comes at the cost of photorealism. The resulting simplistic, composed scenes create a significant domain gap with the complex, organic images used to train modern VLMs. This gap calls into question the evaluation's relevance, as model performance on such synthetic data may not translate to real-world compositional understanding.

Unifying Realism and Concept-driven Generation. Our work is the first to unify these desired properties. We introduce **Auto-Comp**, a framework that pairs the concept-driven control of synthetic benchmarks with the photorealism of data-driven approaches. As summarized in Table 1, our pipeline fully automatically generates high-fidelity images and realistic text from any user-defined concept. Crucially, our framework is the first to introduce the parallel generation of *Minimal* and *Contextual* visio-linguistic conditions for each concept. This parallel construction provides a unique diagnostic tool for measuring the performance gap between a model's foundational binding capacity in sterile environments and its practical robustness in complex, realistic scenes.

3. Methodology

As in Figure 1, the process consists of three main phases.

Concept Definition. The foundation of our pipeline is the *Concept*, a structured definition of a scene's intended ground truth. A concept C is formally defined as a tuple

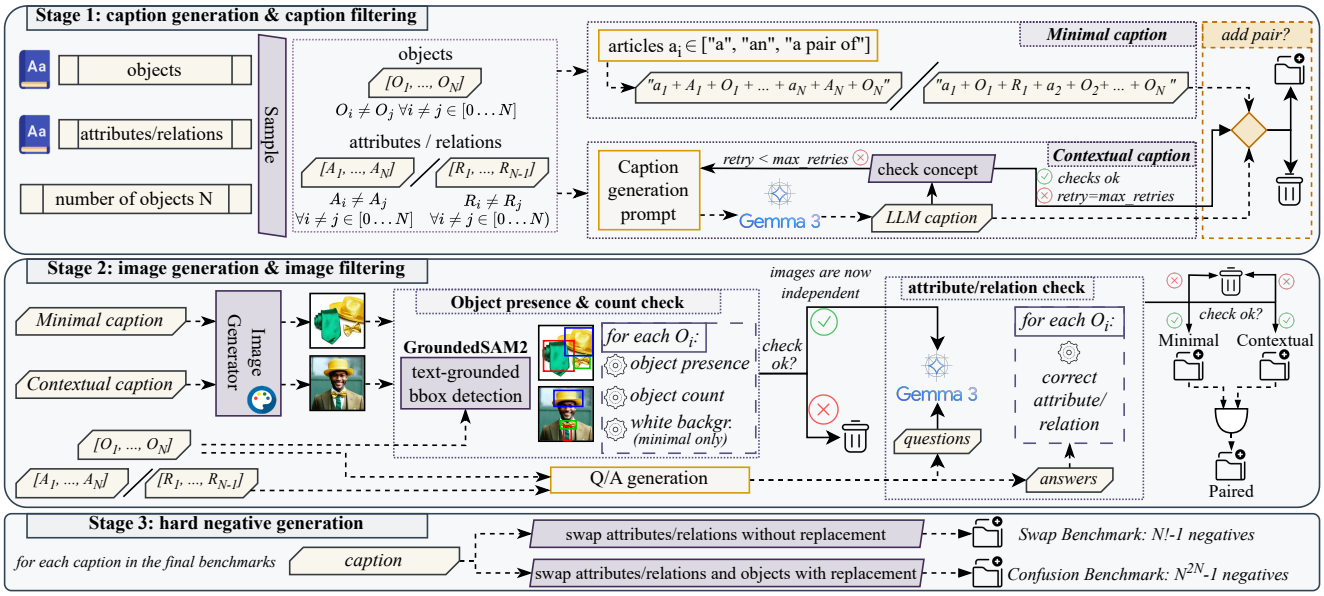

Figure 1. The Auto-Comp pipeline for automated benchmark creation. (1) Caption Generation: Concept-driven generation of template-based Minimal and LLM-based Contextual captions. (2) Image Synthesis & Validation: A text-to-image model generates visuals, which are then validated using GroundingDINO and VLM checks. (3) Hard Negative Generation: Validated pairs are systematically altered by swapping attributes to create challenging compositional test sets.

$C = (\mathcal{O}_C, \mathcal{A}_C)$, where $\mathcal{O}_C = \{o_1, \ldots, o_N\}$ is a set of N objects and $\mathcal{A}_C$ is a set of attributes. The structure of $\mathcal{A}_C$ is task-dependent: for *Color Binding*, it contains N colors with a one-to-one mapping to objects, while for *Position Binding*, it contains $N-1$ binary spatial relations that chain the objects in sequence. The parameter N controls compositional complexity, with $N = 1$ serving as a baseline and $N \geq 2$ used to evaluate true compositional binding. Vocabularies are detailed in Section 4.

Phase 1: Parallel Caption Generation. For each concept, our pipeline generates two distinct styles of captions in parallel. This dual-track generation, designed to represent the same ground-truth with different visio-linguistic properties, is the foundation of our A/B testing framework.

Minimal Captions. This track produces a direct, structured description of the concept using a programmatic, template-based approach. The template joins the object-attribute pairs into a factual list, handling grammatical nuances (e.g., articles, pluralization) and adding the suffix "on a white background". The resulting caption serves to test a model's core binding ability in an isolated, unambiguous setting.

Contextual Captions. This track produces a fluent, naturalistic description of the same concept. The corresponding concept is provided to a generative LLM (Gemma3-12b-it [16]) with a carefully engineered few-shot prompt that defines its role and provides instructions and examples. The full prompt is provided in the appendix. To guarantee the LLM's adherence to these instructions, we perform an automated *semantic preservation check* immediately after generation. This check programmatically parses the LLM's output to verify that all original objects and their assigned attributes from the concept C are present and correctly associated. We define a "correct association" with a strict syntactic pattern:

For Color Binding, given an object-color pair (o_i, a_i), the generated text must contain a substring that matches:

$$\langle \texttt{Article} \rangle \ [\langle \texttt{Word} \rangle] \ a_i \ o_i$$

This pattern ensures the color attribute a_i immediately precedes the object o_i, optionally separated by a single modifier word (e.g., "*an elegant red chair*").

For Position Binding, given a chained relation (o_i, r_i, o_{i+1}), the text must match the pattern:

$$\langle \texttt{Article} \rangle \ [\langle \texttt{Word} \rangle] \ o_i \ r_i \ \langle \texttt{Article} \rangle \ [\langle \texttt{Word} \rangle] \ o_{i+1}$$

This ensures the relation r_k correctly links the two objects, while allowing for an optional modifier word before each.

If the LLM's output fails to match these patterns for all required bindings after several retries, the entire concept is discarded. This means both the failed Contextual Caption and its successfully generated Minimal Caption are removed. This strict pairing protocol guarantees a clean, one-to-one correspondence for every concept that proceeds to the image generation phase.

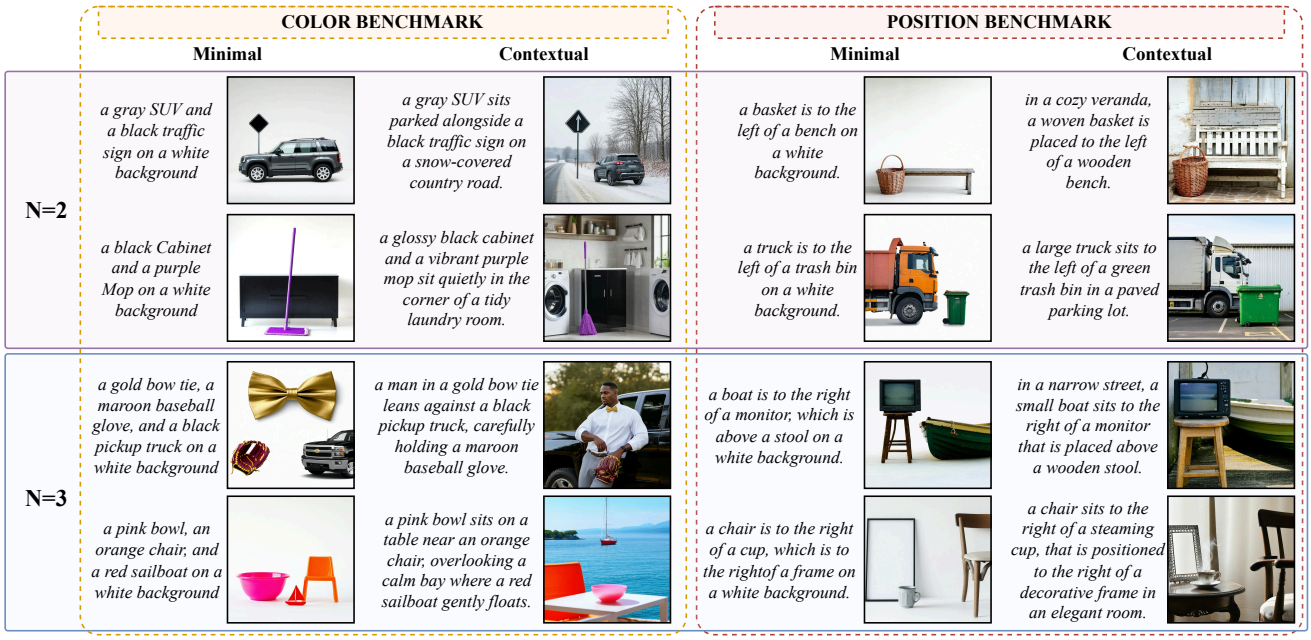

Figure 2. Examples of positive captions-image pairs from the Auto-Comp-CP benchmarks.

Phase 2: Image Synthesis and Automated Validation. From this point forward, the Minimal and Contextual tracks are processed independently to maximize the final dataset sizes. For each caption, we generate an image using StableDiffusion3.5-large [3]. Recognizing that generative models themselves can fail at composition, we introduce a rigorous, two-stage validation pipeline for every generated image to ensure its fidelity.

Object Presence and Count Check. The first, faster validation stage verifies that all objects from the source concept C are present in the image in their correct number. We employ GroundedSAM2 [12, 13], an open-vocabulary detection model, querying it with each object's class name to handle both singular and plural nouns (e.g., ensuring "gloves" corresponds to two instances). Formally, this validation, $V_{\text{obj}}(G(c), C)$, passes only if the detected objects perfectly match the concept's objects $\mathcal{O}_C$ in both class and cardinality.

Background Check for Minimal Images. For images in the Minimal track, we perform an additional validation step. Using the bounding boxes from the object check, we mask all objects and analyze the remaining pixels to verify the background is uniformly white. This check, detailed in the Appendix, allows for minor lighting variations and shadows inherent to the generative process, ensuring each image adheres to the "on a white background" clause and provides a truly sterile visual environment.

Attribute Correctness Check. For an image $I \in \mathcal{I}'_{\text{min}} \cup \mathcal{I}'_{\text{ctx}}$ that has passed the object check, we next verify its attributes. We programmatically generate direct questions based on the source concept C (e.g., "*What color is the monitor?*") and leverage the Gemma-3-12b-it model [16] to provide the answer. To prevent ambiguity, we constrain the VLLM's response to our predefined attribute vocabularies. Formally, we define this validation function as $V_{\text{attr}}(I, C)$, which returns true only if the VLLM provides the correct answer for all attribute questions derived from C.

To ensure the reliability of the proposed image-validation approach, we first conducted a comprehensive human evaluation study (detailed in Section 6) to validate our automated judge. This study established that our VLLM achieves over 94% concordance with expert human annotators, validating its use as a scalable and highly accurate proxy for human judgment.

Positive Benchmark. The positive benchmarks are curated from samples that successfully pass both validation stages. Our large, standalone benchmarks, $\mathcal{B}_{\text{min}}$ and $\mathcal{B}_{\text{ctx}}$, are composed of all validated image-caption pairs for the Minimal and Contextual tracks, respectively. Finally, our *Paired Comparison Set*, $\mathcal{B}_{\text{paired}}$, is created by taking the intersection of these two sets based on their source concepts. This set contains only concepts for which *both* the minimal and contextual versions were successfully validated, enabling a perfectly controlled A/B test. We show qualitative examples in Figure 2. The full benchmark is available in the repository linked in the Abstract and Appendix.

Phase 3: Hard Negative Benchmark Generation The final phase of our pipeline programmatically generates hard negative captions. Unlike prior work that modifies complex,

Table 2. Final number of validated samples in each benchmark.

Benchmark	# Minimal	# Contextual	# Paired
Color N=1	16,325	8,505	4,387
Color N=2	26,388	36,200	15,274
Color N=3	3,074	2,930	1,027
Position N=2	38,393	30,694	17,345
Position N=3	7,478	5,234	1,201
Total	**91,658**	**83,563**	**39,234**

Table 3. Sample survival rates (%) for each benchmark after each validation stage.

	Minimal			Contextual	
Benchmark	**Obj. Check**	**BG Check**	**Attr. Check**	**Obj. Check**	**Attr. Check**
Color N=1	28.6%	24.9%	16.3%	16.6%	8.5%
Color N=2	74.1%	67.6%	26.4%	79.5%	36.2%
Color N=3	33.6%	29.1%	3.1%	35.7%	2.9%
Position N=2	78.8%	65.2%	38.4%	75.4%	30.7%
Position N=3	49.2%	31.6%	7.5%	43.3%	5.2%

real-world captions, our pipeline begins with text whose structure has been programmatically generated and verified. This high degree of control is a key strength. Unlike LLM-based rewriting, which can inadvertently alter sentence structure, our programmatic swapping ensures positive and negative captions are perfectly controlled, identical in length and complexity, differing *only* in the swapped conceptual elements. Our pipeline automatically handles all grammatical details to maintain linguistic naturalness. As we will demonstrate with a blind LLM evaluation (Section 6), this process creates negatives that are free of artifacts, forcing models to reason compositionally. For each concept with $N \geq 2$ objects, we generate two distinct hard negative benchmarks:

The Swap Benchmark. This benchmark is a direct test of compositional binding. Our framework handles two general cases for creating swap-based negatives. For *Attribute Binding Tasks* involving properties like color, we generate negatives by permuting the N attributes among the N fixed objects. For *Relational Binding Tasks* (e.g., spatial position), where permuting the relations is nonsensical for $N = 2$, we instead create foils by permuting the N objects themselves (e.g., "*a table on top of a chair*" from "*a chair on top of a table*"). In either case, this process generates $N! - 1$ unique hard negatives, directly measuring a model's ability to associate each element with its correct role.

The Confusion Benchmark. To move beyond simple binding errors, we design the Confusion Benchmark, a more strenuous test intended to probe for reliance on brittle heuristics and susceptibility to low-entropy distractors. Here, we generate an exhaustive set of captions by creating all possible combinations of objects and attributes from the original concept, *with replacement*. The generation process depends on the attribute type. For an attribute like color, where each of the N objects is assigned one of N colors, we generate all N^N object arrangements and all N^N color arrangements. The Cartesian product yields $N^{2N}-1$ hard-negative captions. For position, which describes the $N-1$ relative relations between the N objects, the calculation is different. We again generate all N^N object arrangements. However, we generate all possible arrangements for the $N-1$ relation slots, sampling from the $N-1$ available position types. This results in $(N-1)^{N-1}$ unique positional arrangements, producing a total of $N^N(N-1)^{N-1}-1$ hard negatives. This benchmark creates a massive set of challenging distractors, allowing us to quantify a model's robustness.

4. The Auto-Comp-CP Benchmark Suite

Vocabularies. Our Auto-Comp-CP (Color-Position) benchmark relies on 3 starting vocabularies.
Object Vocabulary. To ensure diversity and clarity, we curated our object vocabulary from the 365 classes in the Object365 dataset [15]. We filtered out ambiguous concepts, abstract nouns, and non-object categories (e.g., "person," animals) to avoid common model biases, resulting in a final vocabulary of 250 distinct objects.

Color Vocabulary. We defined a vocabulary of 20 common colors, built upon standard VGA colors (e.g., red, blue, green, yellow, magenta) and supplemented with common shades (e.g., gold, brown) to create challenging evaluation samples with similar yet distinct colors.

Spatial Relation Vocabulary. For relational binding, we selected a set of 4 fundamental binary spatial relations inspired by prior work [14]: *over*, *under*, *to the left of*, and *to the right of*.

Dataset Statistics and Analysis. We generate $100k$ pairs of Minimal/Contextual captions per benchmark in Stage 1 (excluding N=1 color benchmark, for which we start from $20k$ as we use it only for ablation). After Stage 2 and 3, our pipeline yields a large-scale benchmark suite, with over 175,000 validated samples generated across the standalone benchmarks (Table 2). The validation survival rates, detailed in Table 3, allow us to pinpoint the specific failure modes of the generative model. While the initial *Object Presence Check* filters many samples and the subsequent *Background Check* further purifies the Minimal set, the most significant drop in survival rate across both tracks consistently occurs at the final *Attribute Correctness Check*. This reveals that true compositional binding is the primary bottleneck even for the generative model. Notably, the model struggles more with multi-object color binding than with spatial relations: this suggests correctly rendering specific attributes is a harder generative challenge than spatial arrangement as complexity increases.

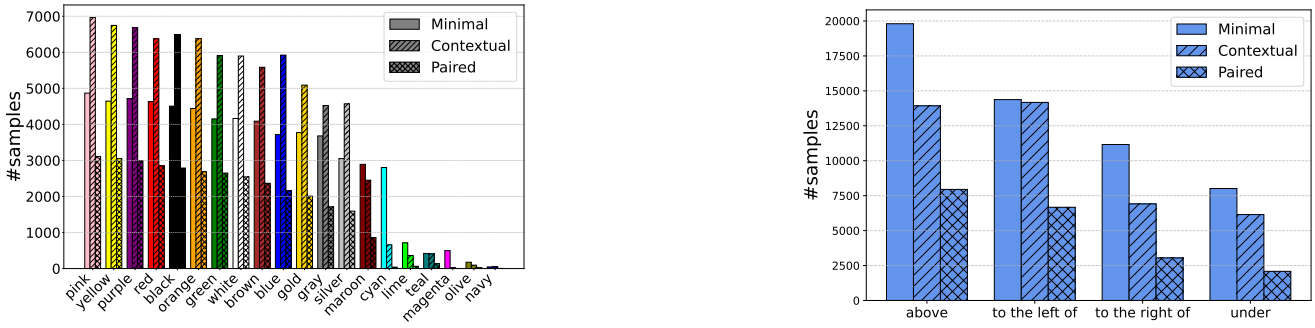

Figure 3. Distribution of validated samples per attribute in our final datasets, summed across N=2 and N=3 tasks. (Left) Total samples per color in the Color Benchmark. (Right) Total samples per relation in the Position Benchmark.

Table 4. Full model performance on Color and Position benchmarks, results in percent (%).
Abbreviations: Min=Minimal, Ctx=Contextual; S=SigLIP, S2=SigLIP2; OC=OpenCLIP; P=Patch size.

		Color Benchmark								Position Benchmark							
		N=2 Objects				N=3 Objects				N=2 Objects				N=3 Objects			
		Min		Ctx		Min		Ctx		Min		Ctx		Min		Ctx	
Family	**Model**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**
	Random Chance	50.0	6.3	50.0	6.3	16.7	0.1	16.7	0.1	50.0	25.0	50.0	25.0	16.7	0.9	16.7	0.9
CLIP	ViT-B-P16	57.3	35.2	61.2	26.7	13.9	2.8	17.4	2.4	53.9	52.5	61.5	55.7	20.1	10.1	27.2	9.9
	ViT-B-P32 (OC)	55.2	34.3	60.4	26.3	15.5	3.0	16.9	2.0	54.3	52.6	62.2	55.8	22.2	9.6	29.1	9.9
	ViT-L-P14 (OC)	54.9	31.2	60.4	25.2	15.0	3.4	15.6	1.8	55.0	54.0	65.2	60.0	21.8	8.3	29.8	9.2
	E02-B-P16 (OC)	61.7	40.0	66.6	29.6	17.2	4.4	17.7	2.4	55.7	54.5	62.7	57.2	22.2	11.2	29.5	9.9
	E02-E-P14 (OC)	65.0	42.4	70.0	38.1	15.5	9.1	20.1	5.6	56.9	56.5	63.7	58.9	22.6	11.0	31.3	10.4
	ViT-H-P14 (OC)	70.6	49.1	74.8	45.4	15.9	11.0	16.7	9.3	58.2	57.9	70.8	67.9	24.0	12.6	36.8	12.8
	ViT-H-P14-378 (OC)	70.4	48.7	75.0	46.4	16.1	11.9	17.7	9.5	58.3	58.1	70.9	67.8	24.0	13.0	37.0	13.6
SigLIP	S-B-P16-224	69.5	43.8	72.3	39.6	17.1	8.0	16.2	4.5	56.8	56.3	71.7	67.7	22.5	11.6	34.3	11.8
	S-B-P16-384	69.3	43.9	72.2	39.7	16.9	7.8	16.2	4.5	56.5	56.1	70.7	66.4	22.9	12.8	34.0	11.6
	S-L-P16-256	73.5	52.0	**76.4**	43.1	16.2	9.2	18.1	6.6	57.9	57.4	75.2	71.0	23.1	10.0	38.6	12.3
	S-L-P16-384	72.7	50.5	75.9	42.8	16.1	8.1	19.9	5.8	57.8	57.3	74.7	70.1	23.1	9.8	38.1	12.4
	S-SO400M-P14-384	71.8	49.9	75.1	42.4	**19.2**	9.4	16.3	6.3	58.3	57.8	75.1	70.7	23.7	11.5	40.5	13.3
	S-SO400M-P14 (OC)	71.6	49.9	74.8	42.5	19.1	9.4	16.2	6.3	58.6	58.1	75.4	70.9	**24.1**	11.6	40.9	13.4
	S2-B-P16-224	73.0	53.7	74.1	42.7	17.4	7.8	15.3	7.5	57.4	57.1	73.4	70.6	21.6	10.2	38.4	13.7
	S2-B-P16-256	72.9	53.7	73.9	42.8	18.3	8.2	15.1	7.4	57.8	57.5	72.9	69.9	21.4	10.4	38.2	13.2
	S2-B-P16-384	72.8	53.7	73.6	42.4	18.6	7.7	14.9	7.4	57.8	57.5	72.4	69.7	21.8	10.5	37.3	13.0
	S2-L-P16-256	73.2	52.9	74.8	43.2	18.0	6.8	18.4	7.3	**59.1**	**59.0**	75.2	72.5	20.7	**13.2**	39.8	16.0
	S2-L-P16-384	72.9	52.9	75.1	**44.0**	16.9	7.0	19.0	8.3	59.0	58.9	75.6	72.8	20.6	13.1	**41.0**	16.3
	S2-SO400M-P14-384	72.0	50.5	75.5	**44.0**	17.6	6.4	17.4	8.7	58.1	57.7	**76.9**	**73.6**	21.3	12.9	40.8	**16.6**
	S2-Giant-P16-256	**74.0**	**54.3**	75.1	43.7	18.8	**11.9**	**20.1**	**9.5**	57.4	57.2	75.9	72.7	20.8	12.3	40.4	15.3

Further analysis of the final attribute distributions (Figure 3), summed over all benchmarks, reveals biases in the generative process. For colors, challenging shades like 'magenta' and 'olive' have lower survival rates, creating a naturalistic long-tail distribution that tests model robustness on both common and rare attributes. For spatial relations, the generator shows a strong preference for 'over', with its counterpart 'under' being the least frequent. This suggests a strong gravitational prior in the generative model, an insight surfaced by our controlled generation process.

5. Experimental Analysis

Main Results: Universal Compositional Failures.

Evaluating 20 VLMs on Auto-Comp-CP reveals universal compositional weaknesses (Table 4). First, performance collapses as complexity increases: while top models reach 74% on Color $N = 2$, they drop to 19% on $N = 3$, barely exceeding random chance. This suggests models rely on simple heuristics rather than a generalizable binding mechanism. Interestingly, models struggle more with Color N=3 than Position N=3, reversing the trend seen at $N = 2$. Second, the Confusion benchmark exposes deeper vulnerabilities than simple attribute swaps. Performance is systematically lower when distractors include repeated attributes or objects (e.g., "red cube and red sphere"). This vulnerability to low-entropy patterns indicates that models are brittle against distractors that humans find trivial, suggesting their reasoning fails when the scene structure exceeds simple permutations. Finally, a training hierarchy emerges: SigLIP models consis-

Table 5. Impact of visio-linguistic context, measured on the Paired Comparison Set. The 'Min' and 'Ctx' columns show 'Swap / Confusion' accuracy (%). The Delta (Δ) columns show the percentage point change for each task. Full results are in the appendix.

	Color Benchmark								Position Benchmark							
	N=2 Objects				N=3 Objects				N=2 Objects				N=3 Objects			
Model	**Min(%)**	**Ctx(%)**	Δ_S	Δ_C	**Min(%)**	**Ctx(%)**	Δ_S	Δ_C	**Min(%)**	**Ctx(%)**	Δ_S	Δ_C	**Min(%)**	**Ctx(%)**	Δ_S	Δ_C
V-H-P14-378	68.2/46.5	67.3/42.2	-0.9	-4.3	**22.0/19.5**	**19.6**/12.2	-2.4	-7.3	60.8/60.7	70.2/67.6	+9.4	+6.9	25.0/12.8	35.9/13.8	+10.9	+1.0
S2-L-P16-384	70.8/51.0	69.2/42.1	-1.6	-8.9	19.1/17.5	17.0/7.2	-2.1	-10.3	61.1/61.0	74.6/72.1	+13.5	+11.1	**25.4/15.4**	41.1/17.2	+15.7	+1.8
S2-Giant-P16	**72.1/52.3**	69.0/41.8	-3.1	-10.5	20.4/16.8	18.3/**9.7**	-2.1	-7.1	**61.4/61.2**	**75.2/72.4**	+13.8	+11.2	24.6/13.4	**42.0/17.7**	+17.4	+4.3

tently outperform standard CLIP variants. The combination of the Sigmoid loss and the WebLI dataset appears crucial for superior compositional ability, as proved by the high performance of even OpenCLIP [5]-trained models that use SigLIP training components.

The Double-Edged Sword of Context Our A/B framework leverages the Paired Comparison Set to isolate the impact of visio-linguistic complexity, comparing sterile Minimal scenes against realistic Contextual ones. The results (Table 5) reveal a surprising, task-dependent trade-off. *Context Aids Spatial Reasoning:* Richer scenes consistently boost performance on the Position task, with gains up to +17.4 points. Realistic environments appear to provide essential geometric and semantic cues that help models disambiguate spatial relationships. *Context Hinders Attribute Binding:* Conversely, added context is universally detrimental to Color binding, causing drops of up to 10.5 points on the Confusion task. This suggests a "visual clutter" effect: background textures and lighting create a noisy signal that overwhelms the model's ability to bind attributes. This highlights a critical brittleness in VLMs: the very context that resolves one form of reasoning severely hinders another.

Deconstructing Failures Beyond Bag-of-Words. The *Confusion Benchmark* enables a deeper analysis of error types, moving beyond simple swaps to probe for more fundamental model biases. The results in Table 6 show that while "swapped colors" is the most frequent error for N=2 compositions (46.9%), the majority of failures stem from other distractors, suggesting the problem is more complex than a bag-of-words limitation. This vulnerability becomes more pronounced at N=3, where the proportion of "swapped colors" errors decreases, while failures due to low-entropy captions with repeated elements rise significantly. The high frequency of these errors demonstrates a strong model bias towards incorrect, simplistic foils, revealing a critical flaw: models not only fail to bind attributes correctly but are also brittle against low-entropy distractors, especially as compositional complexity grows.

6. Analysis and Ablation Studies

Universality of Failures: Generative VLMs and Hard-Negative Miners Evaluating Generative VLMs and Hard-Negative Miners confirms that compositional collapse is

Table 6. Detailed error analysis for the Color benchmark, averaged per modal family.

	N=2 Objects (%)				N=3 Objects (%)			
Family	**Swap Col.**	**Same Col. Diff. Obj.**	**Same Col. Same Obj.**	**Same Obj. Diff. Col.**	**Swap Col.**	**Same Col. Diff. Obj.**	**Same Col. Same Obj.**	**Same Obj. Diff. Col.**
CLIP	44.60	17.72	11.14	26.54	30.96	1.88	3.21	3.14
SigLIP	50.64	12.12	8.94	28.30	38.85	9.61	22.59	28.95
Total	**46.87**	**14.52**	**9.93**	**28.69**	**37.10**	**11.72**	**23.01**	**28.18**

universal (Table 7). While Generative VLMs excel at simple Swaps ($> 88\%$), their failure on the Confusion benchmark proves high-level reasoning cannot compensate for fundamental encoder-level binding flaws. We also replicate the "Double-Edged Sword" trade-off: realistic context aids spatial logic but impairs attribute binding. Finally, Hard-Negative Miners' inability to generalize to Confusion distractors confirms that current training paradigms resolve "bag-of-words" artifacts while remaining vulnerable to the low-entropy binding failures exposed by Auto-Comp.

Single-Object Sanity Check. To confirm that failures are compositional, we tested models on N=1 tasks, where hard negatives are generated by varying either the object or its color with all the possible options in the corresponding vocabularies. As shown in Table 8, top models achieve high accuracy, indicating a solid grasp of individual concepts. As expected, performance is slightly lower in the more challenging *Contextual* setting. Crucially, the dramatic performance drop on $N > 1$ tasks is therefore a specific failure of *binding*, not of basic recognition.

Ruling out Linguistic Biases with a Blind LLM To ensure hard negatives are linguistically indistinguishable from positives, we adopt the "blind LLM" evaluation [4]. We tasked Gemma-3-12b-it with selecting the correct caption from shuffled positive/negative sets without visual input (subsampling 49 negatives for $N = 3$). As shown in Table 9, the LLM performed at near-random chance, confirming our benchmark lacks linguistic artifacts and necessitates genuine visio-linguistic reasoning.

Human Validation Studies. To validate Auto-Comp, four graduate-level evaluators assessed 200 concepts, balanced across all configurations. (1) Concordance: The automated judge achieved $> 94\%$ agreement with humans, confirming its reliability as a ground-truth proxy. (2) Solvability: Humans scored 96% across both Swap and Confusion tasks.

Table 7. Performance of Generative VLMs (top) and Hard-Negative Miners (bottom) on Auto-Comp-CP. Results in percent (%).

		Color Benchmark								Position Benchmark							
		N=2 Objects				N=3 Objects				N=2 Objects				N=3 Objects			
		Min		Ctx		Min		Ctx		Min		Ctx		Min		Ctx	
Family	**Model**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**	**Swap**	**Conf.**
	Random Chance	50.0	6.3	50.0	6.3	16.7	0.1	16.7	0.1	50.0	25.0	50.0	25.0	16.7	0.9	16.7	0.9
Gen VLM	Gemma-3-12B [16]	88.3	71.4	89.4	61.5	84.4	37.2	86.9	21.1	79.0	66.0	89.2	69.4	41.2	23.6	70.6	12.6
	Qwen2.5-VL-7B [1]	89.9	66.0	90.1	61.7	89.8	31.6	89.6	17.7	82.7	69.2	88.7	74.1	55.4	30.7	74.8	19.4
	InternVL3-14B [21]	90.3	74.8	89.7	72.3	82.5	39.8	84.3	34.5	83.2	72.6	87.9	73.9	51.5	35.9	64.4	24.1
Baseline	ViT-B-32 (Base)	55.2	34.3	60.4	26.3	15.5	3.0	16.9	2.0	54.3	52.6	62.2	55.8	22.2	9.6	29.1	9.9
HN Miner	NegCLIP [19]	59.5	35.1	64.0	27.1	18.2	5.5	20.5	4.4	60.1	57.2	66.8	60.5	26.5	10.5	32.6	11.4
	NegCLIP++ [8]	62.8	35.8	66.5	28.8	23.1	5.8	22.3	5.1	63.5	58.1	69.1	64.2	28.4	11.9	37.9	13.1
	TripletCLIP [8]	64.2	37.5	69.1	30.4	23.4	7.2	24.8	6.2	65.2	57.8	72.4	64.1	29.8	12.2	37.5	13.8

Table 8. N=1 benchmark (%).

	Minimal		Contextual	
Model	**Color**	**Obj.**	**Color**	**Obj.**
ViT-H-P14	90.2	93.8	81.7	77.4
S2-Giant-P16	91.7	98.0	85.0	84.5

Table 9. Blind LLM eval (%).

	N=2		N=3	
Model	**Swap**	**Conf.**	**Swap**	**Conf.**
Random	50.0	6.3	16.7	2.0
Gemma3	51.2	6.5	17.4	2.1

Table 10. **Human Validation Studies.** Left: Concordance between our automated pipeline and human judges (%). Right: Human accuracy on the final Swap and Confusion benchmarks (%).

	Minimal		Contextual	
Task	**N=2**	**N=3**	**N=2**	**N=3**
Color	98.0	96.0	96.0	95.0
Position	96.0	94.0	98.0	97.0

		Minimal		Contextual	
Task	**N**	**Swap**	**Conf.**	**Swap**	**Conf.**
Color	2	97.5	97.5	94.5	95.5
	3	97.0	96.0	94.5	94.0
Position	2	97.5	96.5	96.0	95.0
	3	95.5	96.0	93.5	94.0

This stable performance, contrasting with the VLM accuracy drop, proves the "Confusion Gap" is a specific model weakness rather than inherent task ambiguity.

Generalization to New Compositional Concepts A key advantage of the Auto-Comp pipeline is its ability to generalize to diverse compositional phenomena beyond those initially tested. To demonstrate this flexibility, we effortlessly extended our framework to generate benchmarks for two distinct reasoning skills: *Shape-Color Binding* (probing intrinsic object attributes) and *Relative Size* (probing comparative relations). By simply defining new vocabularies, using 3D geometric shapes (e.g., "cube", "pyramid") and size descriptors (e.g., "larger than"), we utilized the unmodified pipeline to produce high-quality Minimal and Contextual datasets. This confirms that Auto-Comp is a general-purpose tool capable of creating targeted probes for user-defined concepts without architectural changes. Crucially, our evaluation of 23 VLMs on these new tasks (detailed in Appendix **??**) replicates the universal failure modes observed in our main analysis, such as the struggle with low-entropy distractors.

Limitations and Bias Mitigation A central challenge in automated benchmarking is the risk of inherited bias: relying on pretrained text-to-image and language models means the resulting data may reflect the priors of these foundation models1. We address this through a "Defense in Depth" strategy designed to isolate and minimize these factors. First, to mitigate generative bias, we employ strict rejection sampling via a multi-stage validation stack, effectively discarding concepts that the model cannot render accurately. We transparently report and discuss the resulting distribution skews due to model biases in 3. Second, to address validator bias, we verified our automated pipeline against human annotators, achieving $> 94\%$ agreement. Finally, our Minimal vs. Contextual A/B design provides structural control, allowing us to distinguish failures caused by core binding issues from those induced by biases in the LLM linguistic style. However, our framework remains subject to the constraints of current generative technology. Since the Auto-Comp pipeline relies on a pretrained tti model, our benchmarks inevitably inherit specific biases that validation cannot fully erase. For instance, our generator exhibited a strong prior for the spatial relation 'over' while struggling with 'under', and showed higher failure rates with rare colors (Figure 3). Similarly, while our automated validation is highly consistent with human judgment, this approach is not infallible and has its own internal biases. Consequently, improving the semantic fidelity of generators and the robustness of model-based judges represents a key area for future work.

7. Conclusion

We introduced Auto-Comp, a fully automated pipeline to generate controllable benchmarks for compositional failures in VLMs. Using our generated Auto-Comp-CP benchmark, we found universal binding failures across 20 models and discovered a surprising trade-off: visio-linguistic context aids spatial reasoning but hinders color binding. We release our pipeline to drive the development and evaluation of more robust VLMs that truly understand complex concepts.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8

[2] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018, 2024. 1, 2

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4

[4] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023. 1, 2, 7

[5] 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, 2021. If you use this software, please cite it as below. 7

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

[7] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 1, 2

[8] Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, et al. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760, 2024. 8

[9] 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*, pages 8748–8763. PmLR, 2021. 1, 2

[10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022.

[11] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022. 1

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 4

[13] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 4

[14] Massimo Rizzoli, Simone Alghisi, Olha Khomyn, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. Civet: Systematic evaluation of understanding in vlms. *arXiv preprint arXiv:2506.05146*, 2025. 1, 2, 5

[15] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 5

[16] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid,

Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. 3, 4, 8

[17] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[18] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2

[19] 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 *The Eleventh International Conference on Learning Representations*. 8

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[21] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, 2025. 8