Survey on Compositionality in VLMs: A Review of Recent Advances

Rozhan Ahmadi *roz.ahmadi@sharif.edu*
Department of Computer Engineering
Sharif University of Technology

Mohammad Mohammad Beigi *moh.moh@sharif.edu*
Department of Computer Engineering
Sharif University of Technology

Parham Moghaddasi *parham.moghaddasi04@sharif.edu*
Department of Computer Engineering
Sharif University of Technology

Mohammadmostafa Rostamkhani *mohammadmostafarostamkhani@gmail.com*
Department of Computer Engineering
Sharif University of Technology

Abstract

Vision–language models (VLMs) such as CLIP achieve strong zero-shot classification and retrieval performance, yet a growing body of work shows that this performance coexists with a severe inability to reason compositionally: to bind attributes to the objects they modify, to resolve relations between entities, and to treat the order of words and objects as semantically meaningful rather than as an unordered bag of concepts. This survey reviews the recent literature on compositionality in VLMs, organizing the field's impactful works into three non-mutually-exclusive categories: *benchmarks and evaluations*, which build the datasets and metrics that make compositional failure measurable; *analyses*, which put forward theories, hypotheses, and empirical investigations into what VLMs are missing and why they behave as they do; and *mitigations*, which propose methods to improve compositional performance or to repair the specific weaknesses that this literature has identified. Each category is further divided into sub-categories that group works by the particular problem they address, so that the reviewed papers are read as a connected line of argument rather than as independent contributions. Building on this organization, the survey then compares the reviewed works against one another along the dimensions that only become visible when they are read together, and closes by identifying the open gaps this comparison exposes together with the research directions they motivate.

1 Introduction

The capacity to understand structured combinations of concepts, referred to as *compositionality* in linguistics, is a defining feature of human language and reasoning: two descriptions built from the same set of words can pick out fundamentally different scenes, and while resolving this is effortless for humans, a growing body of work shows it to be a much harder problem for vision–language models (VLMs).

Foundational VLMs such as CLIP (Radford et al., 2021) have achieved strong results on zero-shot classification and image–text retrieval, and are increasingly embedded in downstream applications, from image generation guided by text prompts to embodied agents following natural-language instructions. High scores on these standard evaluations can nonetheless coexist with deep failures in compositional reasoning. As VLMs are deployed in settings where correctly interpreting compositionally complex descriptions is critical to safe and reliable behavior — where the difference between an instruction to place one object on top of another and its reverse is the difference between success and failure — understanding, detecting, and mitigating these failures becomes an increasingly urgent problem.

Compositionality in VLMs refers to a model's ability to systematically combine simpler concepts, such as objects, attributes, relations, and spatial arrangements, in order to reason about novel and more complex descriptions, going beyond memorized associations or keyword matching. A model may correctly associate an image with a caption not because it has understood the relational structure of the scene, but because the two share the same collection of concepts; this shortcut works well in standard retrieval and breaks down as soon as fine-grained structural distinctions are required. Compositional failures in VLMs tend to cluster around a few recurring patterns: incorrectly assigning an attribute to the wrong object (*attribute binding* failure), failing to reason about spatial or directional relations between objects (*relational* failure), and failing to generalize beyond learned statistical associations to unseen combinations of familiar concepts (concept association and generalization failure).

Recognizing the importance of compositional reasoning in VLMs, a growing body of work has investigated these shortcomings from several angles: constructing benchmarks on which a model cannot succeed by recognizing concepts alone, analyzing why the failures arise, and proposing strategies to mitigate them. This survey reviews that literature and, more importantly, reads it against itself. Taken individually the papers tell a consistent story; compared side by side they disagree about how large the compositional deficit is, about where in the model it originates, and about whether the interventions proposed to close it have in fact done so. Surfacing these disagreements, and identifying what would be required to settle them, is the contribution this survey aims to make beyond summarizing the individual works.

Scope. This survey covers compositional *understanding* in image–text models: given an image and a description, whether the model respects the structure of that description rather than merely recognizing the concepts it contains. Both major architectural families fall within this boundary, dual-encoder models such as CLIP (Radford et al., 2021) and generative VLMs that couple a vision encoder to a language model, since the reviewed literature evaluates and intervenes on both, and several of its central findings concern precisely how the two differ. Three categories of contributions are reviewed. Benchmarks and evaluations are included because studying compositional ability requires measures that distinguish it from ordinary concept matching, and because the validity of these measures remains a matter of debate. Analyses are covered because the same observed failure admits several competing explanations, located variously in the training objective, the architecture, the data, and the evaluation procedure. Mitigations are covered because they constitute the field's practical response and, when compared against one another, reveal which explanations have proved actionable. The period covered runs from 2022 — when Thrush et al. (2022) and Yuksekgonul et al. (2022) established compositional failure as a systematic phenomenon — to the present. Several related lines of work are deliberately set aside to keep this focus: compositional generalization in text-only language models, the compositionality of text-to-image *generation*, video–language and embodied settings whose temporal and action-level structure raises separate questions, and compositional zero-shot learning except where it is used to probe the structure of pretrained representations (Berasi et al., 2025).

Organization. Section 2 introduces the background concepts, notation, and evaluation preliminaries needed to follow the rest of the survey. Section 3 describes how the reviewed papers were collected and organized, and states the principles behind the taxonomy. Following that taxonomy, Section 4 reviews the papers in each category. Section 5 compares the reviewed works along dimensions that only become visible when they are read together, and Section 6 identifies the open problems these comparisons expose together with the research directions they motivate.

2 Background

2.1 Vision–Language Models and Notation

Throughout, I is written for an image, C for a caption, and $s(I, C) \in \mathbb{R}$ for the score a model assigns to their pairing. The VLMs relevant to this survey fall broadly into two architectural families, which differ in how s is computed and therefore in how compositional competence can be probed.

Dual-encoder (or dual-stream) models, exemplified by CLIP (Radford et al., 2021), encode images and text independently with separate towers, $v = f_{\text{img}}(I)$ and $t = f_{\text{txt}}(C)$, and align them in a shared embedding space through a contrastive objective, so that matching pairs are pulled together and mismatched pairs are pushed apart. For a batch of N pairs with cosine similarity $s(v, t)$ and temperature τ, the objective is

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{\exp\big(s(v_i, t_i)/\tau\big)}{\sum_{j=1}^{N} \exp\big(s(v_i, t_j)/\tau\big)} + \log \frac{\exp\big(s(v_i, t_i)/\tau\big)}{\sum_{j=1}^{N} \exp\big(s(v_j, t_i)/\tau\big)} \right].$$

Because similarity is computed from two independently pooled vectors, dual-encoder models are cheap to use for large-scale retrieval; as discussed throughout this survey, that same pooling step is repeatedly identified as a bottleneck for encoding structural information (Kamath et al., 2023b; Oh et al., 2024). Note also that the negatives in the denominator are the other captions in a randomly sampled batch, which typically differ from C_i in many respects at once — a property that Section 2.3 shows to be central to the problem.

Generative VLMs (GVLMs) instead connect a visual encoder to a large language model and produce text autoregressively conditioned on the image, as in LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023), and BLIP-2 (Li et al., 2023). Compositional evaluation of such models typically scores a candidate caption by its log-likelihood,

$$s(I, C) \;=\; \sum_{k=1}^{|C|} \log P\big(c_k \mid c_{<k},\, I;\, \theta\big),$$

which, as Ma et al. (2024) show, introduces a failure mode of its own: the language-modeling component can inflate the score of a fluent caption independently of its visual relevance.

Both families are adapted to downstream or compositional objectives through a range of fine-tuning paradigms discussed in this survey: full-parameter contrastive fine-tuning with hard negatives (e.g. NegCLIP (Yuksekgonul et al., 2022)), parameter-efficient fine-tuning such as LoRA (Hu et al., 2022) or QLoRA (Dettmers et al., 2023) adapters applied to GVLMs (Awal et al., 2024; Mishra et al., 2025), fine-tuning of a single modality tower while the other is frozen (e.g. text-encoder-only updates in Peleg et al. (2025)), and lightweight auxiliary modules — cross-modal attention layers (Tang et al., 2023) or dependency-guided decoders (Parascandolo et al., 2025) — attached to an otherwise frozen backbone. Section 4 discusses how these paradigms interact with the specific failure modes they target.

2.2 Compositionality in the Multimodal Setting

In linguistics, compositionality refers to the principle that the meaning of a complex expression is determined by the meanings of its constituents and the rules used to combine them. Applied to VLMs, this becomes an operational question: can the model distinguish two image–caption pairs that share the same concepts but differ in their structural arrangement? Three types of compositional competence recur across the surveyed literature. *Attribute–object binding* is the ability to link a property such as color or size to the specific object it modifies rather than to the scene as a whole. *Relational understanding* is the ability to encode directional or structural relationships between entities, distinguishing "the dog is chasing the cat" from "the cat is chasing the dog". *Order and structure sensitivity* is the ability to treat the arrangement of words as semantically meaningful rather than as an interchangeable collection of tokens. Later benchmarks add axes to this list — negation, counting, multi-hop spatial chaining, and the interpretation of lexically fused expressions — but these three define the problem as it was originally posed.

2.3 The Bag-of-Words / Bag-of-Concepts Problem

The failure mode that motivates most of this literature is the tendency of VLMs, particularly those trained contrastively, to behave as though a caption were an unordered set of concepts. The behavior can be stated precisely. Suppose the score a model assigns decomposes, at least approximately, as a sum of independent concept-level contributions,

$$s(I, C) \approx \sum_{w \in C} \varphi(w, I),$$

where φ measures the presence of concept w in image I. Then for any permutation π of the caption's tokens, $s(I, C) = s(I, \pi(C))$: the model is constitutively incapable of distinguishing a caption from its reordering, no matter how much the reordering changes the meaning.

The reason such a scoring function is learned is economic rather than architectural. Given web-scale training data and in-batch negatives drawn at random, the candidate captions competing with the correct one typically differ from it in several concepts simultaneously, so recognizing *which* concepts appear is already sufficient to minimize the contrastive loss. Encoding *how* those concepts are arranged supplies little additional discriminative signal and is therefore not learned. Bag-of-words behavior is, on this account, the rational solution to the objective; the mitigation literature of Section 4.3 can largely be read as a sequence of attempts to change the objective so that it stops being rational.

2.4 Hard Negatives and Hard Positives

The dominant intervention in this literature follows directly from the argument above. A *hard negative* is a caption (or image) constructed to share nearly all concepts with the correct one while differing in structure — a swapped relation, a reassigned attribute, a permuted word order — so that concept matching alone cannot separate the two. Adding a set $\mathcal{N}(C_i)$ of such negatives to the denominator of the contrastive loss yields the objective used, in variations, by most methods surveyed here.

A *hard positive*, correspondingly, is a caption whose surface form differs from the original while its meaning is preserved — a paraphrase or synonym substitution — and which should therefore retain a high score. The distinction matters because the two are not symmetric in effect: as Kamath et al. (2023a) show, training exclusively against hard negatives teaches a model to penalize surface change in general, including changes that preserve meaning. Much of the analysis in Section 4.2 concerns the consequences of this asymmetry, and how the quality of $\mathcal{N}$ determines whether the resulting gains are genuine.

2.5 Baselines: Chance and Human Performance

Because the protocols defined in the literature differ in how a correct answer is defined, an accuracy figure is uninterpretable on its own. Papers therefore report results against two reference points: *random chance*, which fixes the floor, and *human performance*, which estimates the ceiling. The distance between them is the headroom a benchmark actually measures, and comparing raw scores across benchmarks without accounting for both is a persistent source of confusion.

2.6 Common Data Sources and Backbones

A final piece of shared context is that most of this literature is built on a small set of resources, which contributes to the correlations and complications discussed later. Benchmarks and training sets are overwhelmingly derived from COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and Flickr30k (Plummer et al., 2015), with LAION (Plummer et al., 2015) and CC3M/CC12M (Changpinyo et al., 2021) used for pretraining-scale work; hard negatives are generated by LLMs and hard-negative images by diffusion models, so that the properties of those generators enter the evaluation pipeline. On the model side, the contrastive results reported throughout concern CLIP and its close relatives (SigLIP (Zhai et al., 2023), BLIP (Li et al., 2022), FLAVA (Singh et al., 2022), CoCa (Yu et al., 2022), XVLM (Zeng et al., 2021)), while generative results concern LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023), BLIP-2 (Li et al., 2023), and Idefics2 (Laurençon et al., 2024), alongside closed-source systems used as points of comparison. The

concentration is worth keeping in view: conclusions in this field are, to a first approximation, conclusions about CLIP-family models evaluated on COCO-derived data.

3 Methodology and Taxonomy

3.1 Paper Selection

The survey covers thirty-one papers published between 2022 and 2026 (present), comprising the foundational works in the field together with those that have most shaped its subsequent direction. The starting points are the two studies that established compositional failure as a systematic phenomenon rather than an isolated observation (Thrush et al., 2022; Yuksekgonul et al., 2022); the remaining works are those that responded to them, whether by measuring the phenomenon more precisely, by explaining it, or by attempting to repair it. Three criteria govern inclusion. A paper must make a substantive contribution to compositional understanding in image–text models — a benchmark, a diagnostic result, or an intervention — rather than reporting compositional performance incidentally. Its claims must be evaluated on at least one benchmark shared with other work in the set, which is what makes the comparisons of Section 5 possible at all. And it must fall within the boundaries set out in the scope paragraph of Section 1.

3.2 Organizing Principle

Papers are organized by their *primary contribution* into three categories — benchmarks and evaluations, analyses, and mitigations. Two properties of this scheme are worth stating explicitly.

The categories are *non-mutually-exclusive*. A single paper frequently introduces a benchmark, uses it to diagnose a failure, and proposes an intervention derived from that diagnosis; of the thirty-one papers surveyed, twelve contribute to all three categories and only five to exactly one, as Table 4 records. Forcing such works into a single bucket would misrepresent them and, more importantly, would obscure the loop — measure, explain, repair — that turns out to be the field's characteristic unit of contribution. Each paper is therefore discussed in every subsection to which it contributes, with cross-references making the connections explicit.

Within each category, papers are grouped into second-level clusters ordered by *the problem each cluster addresses*, since each wave of work responds to a limitation exposed by the wave before it. Benchmarks become more tightly controlled as earlier benchmarks are shown to be exploitable; analyses relocate the explanation for failure as each account is shown to be partial; and mitigations move from the training data to the loss, the objective, the architecture, and finally to inference alone as the cost of each intervention becomes apparent.

3.3 Second-Level Structure

The benchmark category divides into seven clusters. *Foundational probes* establish the phenomenon; *component-level diagnostics* attribute it to a particular stage of the pipeline; *benchmark-artifact correction* repairs evaluations shown to be exploitable; *visual and bidirectional evaluation* extends measurement to perturbed images and to the text-to-image direction; *beyond binary retrieval* covers formats that abandon the two-candidate decision; *new compositional axes* opens dimensions the object–attribute–relation template does not reach; and *scoring procedures* revises how predictions are converted into a score.

The analysis category divides into eight clusters, each locating the cause of failure in a different place: *contrastive learning as a rational shortcut*, *architectural bottlenecks*, *evaluation-layer Problems*, *mitigation side effects*, *training objectives and learning dynamics*, *recognition without structure*, *capability-level bottlenecks*, and *non-linear compositional geometry*.

The mitigation category divides into seven clusters by where the intervention acts: *foundational hard-negative interventions*, *capability-preserving fine-tuning*, *visual hard negatives*, *redesigning the training objective*, *axis-specific interventions*, *architectural and data-centric recommendations*, and *inference-time methods*. Figure 1 presents the full taxonomy; Tables 1, 2, and 3 summarize each category's clusters with their representative works.

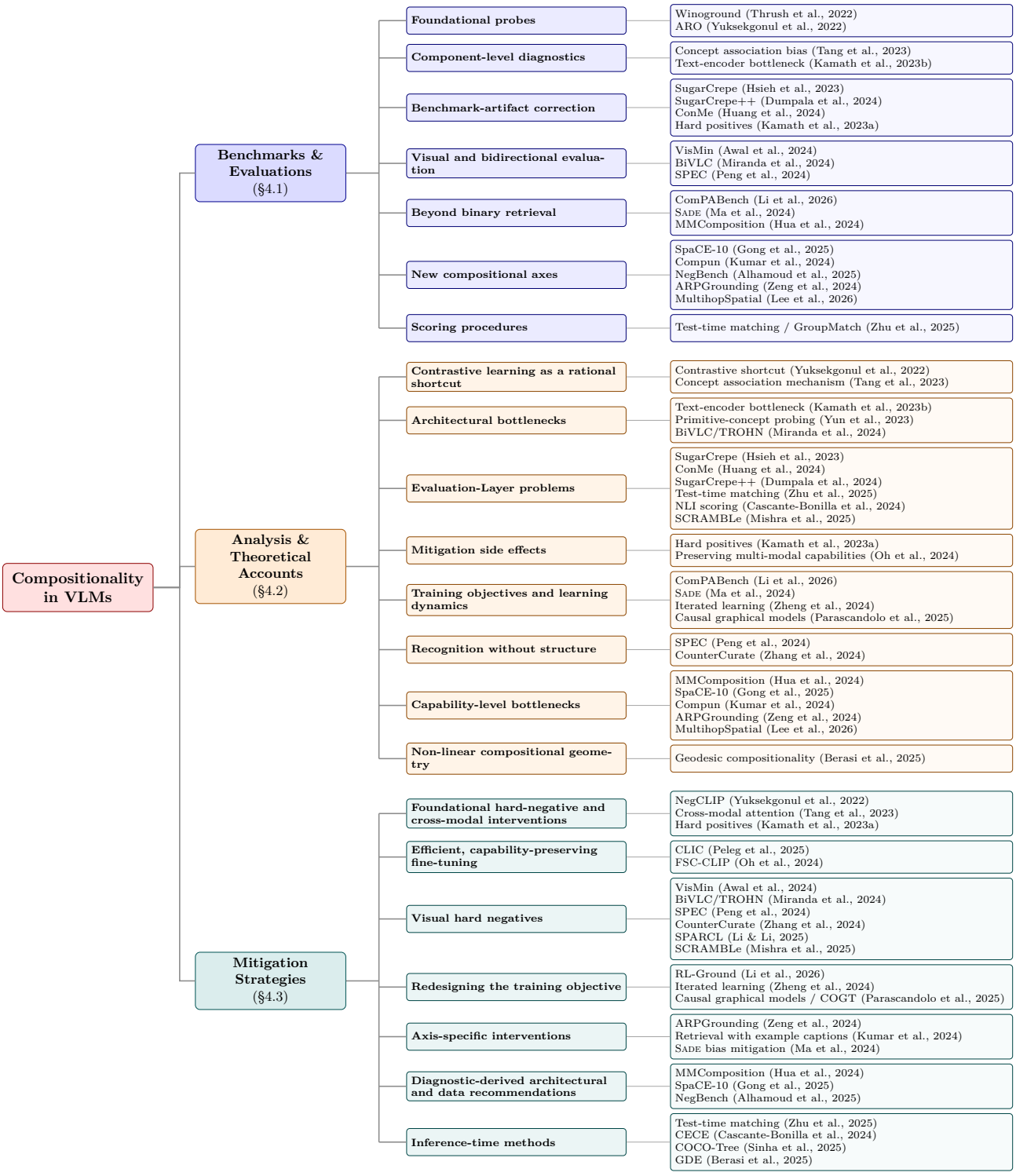


Figure 1: Taxonomy of the reviewed literature. The three branches correspond to the primary-contribution categories of Section 3; each branch expands into its second-level clusters, ordered by the problem each addresses rather than by publication date, and each cluster expands in turn into the papers assigned to it. The categories are non-mutually-exclusive: of the thirty-one papers surveyed, eleven appear under all three branches — introducing a benchmark, diagnosing the failure it exposes, and proposing an intervention derived from that diagnosis — while seven appear under exactly one. Tables 1, 2 and 3 give the same mapping in tabular form.

Category	Representative papers
Foundational probes	Winoground (Thrush et al., 2022) ARO (Yuksekgonul et al., 2022)
Component-level diagnostics	Concept association bias (Tang et al., 2023) Text-encoder bottleneck (Kamath et al., 2023b)
Benchmark-artifact correction	SugarCrepe (Hsieh et al., 2023) SugarCrepe++ (Dumpala et al., 2024) ConMe (Huang et al., 2024) Hard positives (Kamath et al., 2023a)
Visual and bidirectional evaluation	VisMin (Awal et al., 2024) BiVLC (Miranda et al., 2024) SPEC (Peng et al., 2024)
Beyond binary retrieval	ComPABench (Li et al., 2026) SADE (Ma et al., 2024) MMComposition (Hua et al., 2024)
New compositional axes	SpaCE-10 (Gong et al., 2025) Compun (Kumar et al., 2024) NegBench (Alhamoud et al., 2025) ARPGrounding (Zeng et al., 2024) MultihopSpatial (Lee et al., 2026)
Scoring procedures	Test-time matching / GroupMatch (Zhu et al., 2025)

Table 1: Summary of benchmark and evaluation papers reviewed in Section 4.1.

4 Review of Key Papers

Compositional understanding is fundamental to most tasks that vision–language models are expected to perform. A retrieval system that cannot tell which object an attribute modifies will return the wrong image whenever a query contains two objects; a generation pipeline conditioned on a compositionally insensitive text encoder inherits that insensitivity; an agent that cannot distinguish a relation from its converse will act on the wrong one. Rather than being a curiosity at the margins, such failures constitute a defining bottleneck of the paradigm, which is why the past four years have produced a substantial literature devoted to measuring, explaining, and repairing them.

This section reviews the key works on compositionality in VLMs, following the taxonomy introduced in Section 3. *Benchmarks and evaluations* (Section 4.1) construct the data and metrics that make compositional gaps measurable, and are reviewed across seven clusters. *Analyses* (Section 4.2) ask why the failures occur, and are reviewed across eight clusters, each locating the cause in a different part of the pipeline. *Mitigations* (Section 4.3) intervene to close the gap, and are reviewed across seven clusters ordered by where the intervention acts.

4.1 Benchmarks & Evaluations: Measuring Compositional Gaps

A fundamental challenge in measuring compositional understanding is that the tasks vision–language models are conventionally evaluated on do not require it. Standard image–text retrieval rewards a model for recognizing which concepts a caption mentions; because the competing captions in a randomly sampled batch differ from the target in many ways at once, whether those concepts are arranged as the caption specifies rarely changes the ranking. Strong retrieval scores are therefore fully compatible with complete insensitivity

to structure, and the benchmarks reviewed here exist to remove that compatibility by constructing candidate sets in which arrangement is the only thing that distinguishes the correct answer from the incorrect one.

Two distinct contributions are grouped together in this section. A *benchmark* supplies the data: a set of items, each pairing an image or images with candidate descriptions, constructed so that concept overlap alone cannot identify the correct pairing. An *evaluation* supplies the procedure: the rule that converts a model's scores over those candidates into a judgment of success, together with the baseline against which that judgment is read (Section 2.5). The two are separable, and the distinction matters, since a benchmark can be revised while its scoring rule is retained, and a scoring rule can be revised without touching the data at all.

Within these two contributions, the works reviewed here differ along a small number of design dimensions that determine what each is able to detect. A benchmark must decide which compositional phenomenon it targets, whether attribute binding, relational structure, word order, or any other axis such as counting or negation. It must decide how its incorrect candidates are produced — by rule, by a language model, by an image-conditioned model, or by hand — since this determines whether they can be rejected for reasons other than the one under test. It must decide which side is perturbed, the caption or the image, and therefore which retrieval direction is exercised. And it must decide what a model is asked to do with the candidates: choose between two, choose among many, accept and reject simultaneously, or point to a region of the image. The clusters that follow group the reviewed works by the design decision each was introduced. Table 1 summarizes this taxonomy.

4.1.1 Foundational Probes

This group covers the first benchmarks built specifically to test whether vision–language models are sensitive to structure. Their aim was narrow: to show that the problem exists and is widespread, not yet to explain it. Both do so by holding the words of a caption fixed and changing only their arrangement, so that a model matching concepts alone has nothing left to rely on. They differ mainly in how they were built, one written by hand at small scale and one generated automatically at large scale.

The founding evaluations were designed around a deliberately narrow question: does the order of words in a caption make any difference to a vision–language model? **Winoground** (Thrush et al., 2022) poses that question in its purest form. Each of its 400 manually curated items contains two images and two captions built from an identical set of words arranged differently—"some plants surrounding a lightbulb" against "a lightbulb surrounding some plants" (Figure 2)—so that lexical overlap carries no information whatsoever and only the resolution of compositional structure can separate the pairs. Scoring is correspondingly strict: alongside a Text Score and an Image Score for each retrieval direction, the Group Score credits a model only when it aligns both images and both captions correctly at the same time.

$$\text{Text Score} = f(C_0, I_0, C_1, I_1) = \begin{cases} 1, & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \text{and } s(C_1, I_1) > s(C_0, I_1), \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Image Score} = g(C_0, I_0, C_1, I_1) = \begin{cases} 1, & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \text{and } s(C_1, I_1) > s(C_1, I_0), \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Group Score} = h(C_0, I_0, C_1, I_1) = \begin{cases} 1, & \text{if } f(C_0, I_0, C_1, I_1) \\ & \text{and } g(C_0, I_0, C_1, I_1), \\ 0, & \text{otherwise.} \end{cases}$$

The design that gives Winoground its precision also bounds it. As Diwan et al. (2022) observe, the benchmark is small, and many of its items additionally demand commonsense inference, world knowledge, or pragmatic interpretation, so a failure cannot be attributed to compositional structure alone.

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

Figure 2: An example from Winoground. The two sentences contain the same words but in a different order. The task of understanding which image and caption match is trivial for humans but much harder for vision and language models. Adapted from Thrush et al. (2022).

To address the Winoground's limited scale, **ARO** (Attribution, Relation, and Order) (Yuksekgonul et al., 2022) answers generates hard negatives programmatically rather than curating them by hand. Its four subsets, together comprising over 50,000 examples, perturb relations, attribute bindings, and word order through rule-based operations such as swapping the nouns flanking a relation or shuffling the tokens of a caption, allowing the phenomenon Winoground had demonstrated on a handful of items to be probed across standard captioning corpora. The trade-off inherent in the approach is precisely what motivates the next group of benchmarks: a rule that reliably alters meaning offers no guarantee of preserving plausibility or grammaticality.

4.1.2 Component-Level Diagnostics

Showing that models fail at compositional tasks does not say where the failure enters the pipeline. A single benchmark score mixes together everything the model does, so a low result is equally consistent with a visual problem, a linguistic one, or a breakdown in how the two are matched. The two benchmarks in this group are built to separate these possibilities: each holds one part of the process fixed while varying another, so that a failure can be traced to a particular stage rather than to the model as a whole.

The first examines whether learned co-occurrence statistics override visual evidence. Tang et al. (2023) construct controlled evaluations using the **Natural-Color Dataset (NCD)**, **UNnatural-Color Dataset (UNCD)** of Anwar et al. (2025), and Rel3D-based modifications from Goyal et al. (2017), where objects are paired with attributes that violate their typical associations. By comparing prior-consistent and prior-violating examples while holding other factors fixed, the benchmark isolates the effect of learned concept associations and reveals when models rely on pretraining statistics rather than the visual input itself.

The second localizes failure within the vision-language architecture. Kamath et al. (2023b) pair a text-only benchmark, **CompPrompts**, with a matched multimodal counterpart, **ControlledImCaps**, evaluating the same phenomena—including spatial and temporal relations, negation, and attribute binding—with and without visual input. If a model already fails in the text-only setting, the missing information likely originates before cross-modal alignment, pointing to limitations in textual representation rather than multimodal fusion. Beyond diagnosis, this framework provides a principled basis for deciding whether future interventions should target the text encoder, vision encoder, or alignment mechanism.

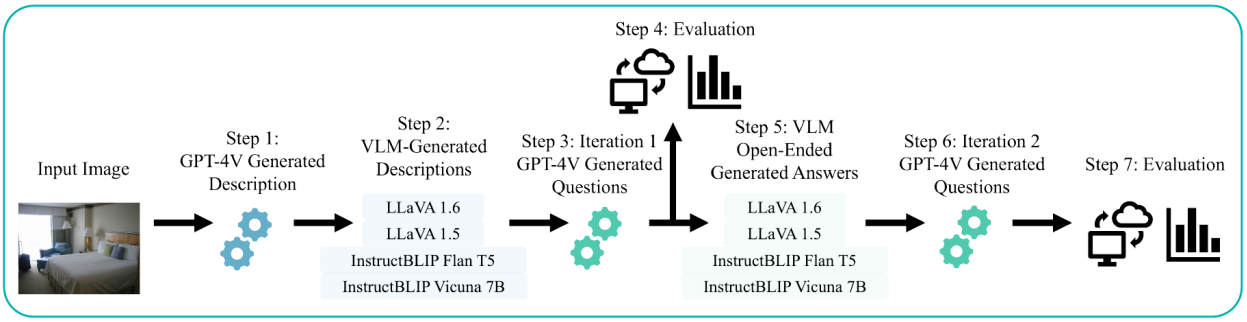


Figure 3: Huang et al. (2024)'s proposed CR data generation framework employs multiple open VLMs in a multi-stage 'conversation' setup. Adapted from Huang et al. (2024).

4.1.3 Benchmark-Artifact Correction

As rule-based compositional benchmarks became common, attention turned to whether the benchmarks themselves were introducing shortcuts. If an incorrect caption can be rejected because it is ungrammatical, describes an impossible scene, or is otherwise odd, then a model can score well without ever consulting the image, and the measured gap reflects how the benchmark was built as much as what the model can do. The benchmarks in this group exist to repair earlier ones. Each removes a shortcut that the previous correction had left in place.

SugarCrepe (Hsieh et al., 2023) provides the first correction by replacing mechanically generated negatives with fluent, plausible alternatives produced through LLM-based editing. Candidate negatives are filtered to remove examples that a text-only classifier can solve without visual information, reducing reliance on superficial linguistic cues and narrowing the measured compositional gap.

SugarCrepe++ (Dumpala et al., 2024) identifies a remaining shortcut in the evaluation format itself. Binary discrimination between a positive and a negative caption cannot distinguish semantic understanding from sensitivity to surface-level changes. The benchmark therefore introduces a three-way formulation containing the original caption, a meaning-preserving paraphrase, and a meaning-altering negative, requiring models to recognize both semantic equivalence and semantic difference.

A further limitation concerns the generation process itself: negatives written without observing the image may be rejected because they are generally implausible rather than because they contradict the specific scene. **ConMe** (Huang et al., 2024) addresses this issue by grounding negative generation in the image through a multi-turn process involving GPT-4V (Yang et al., 2023) and VLM evaluators, producing distractors that are plausible for the depicted scene while differing compositionally from the target caption.

Where these three benchmarks refine the negative, Kamath et al. (2023a) address the neglected positive side of the same problem: whether a model can *accept* a caption whose surface form has changed while its meaning has not. They construct a large-scale hard-positive set from Visual Genome and introduce two metrics to score it, **Augmented Test Accuracy (ATA)**, which requires a model to handle hard positives and hard negatives together, and **Brittleness**, which measures the extent to which rejection behavior spills over onto semantically equivalent variants. Compositional evaluation thereby becomes explicitly two-sided, demanding rejection of altered meanings and acceptance of preserved ones under a single criterion.

$$\text{Augmented Test Accuracy (ATA)} = s(c|i) > s(c_n|i) \text{and } s(c_p|i) > s(c_n|i).$$

4.1.4 Visual and Bidirectional Evaluation

Most compositional benchmarks perturb the caption, which treats compositionality as a question about image-to-text retrieval and leaves the opposite direction untested. The reason is practical: rewriting a sentence to change one relation is easy, while editing a photograph to change one relation and nothing else is not. The benchmarks in this group take on that difficulty, building images that differ from the original

in a single controlled respect. Doing so makes it possible to ask whether a model that handles one retrieval direction handles the other as well.

VisMin (Awal et al., 2024) constructs minimally different image pairs through an LLM-guided editing pipeline with human verification, targeting changes in object identity, attributes, count, and spatial relations (Figure 4). By holding the caption fixed while varying the image along a single dimension, it directly evaluates text-to-image retrieval. **BiVLC** (Miranda et al., 2024) extends this idea by converting SugarCrepe hard-negative captions into corresponding candidate images and evaluating both retrieval directions using Winoground-style metrics (Figure 5), revealing whether improvements from text-side training transfer to visual retrieval.

SPEC (Peng et al., 2024) further increases control by replacing natural scenes with synthetic candidate sets in which examples differ along exactly one property—such as size, existence, position, or count. This design enables failures to be localized to specific structural capabilities rather than reported only as an overall compositionality score.

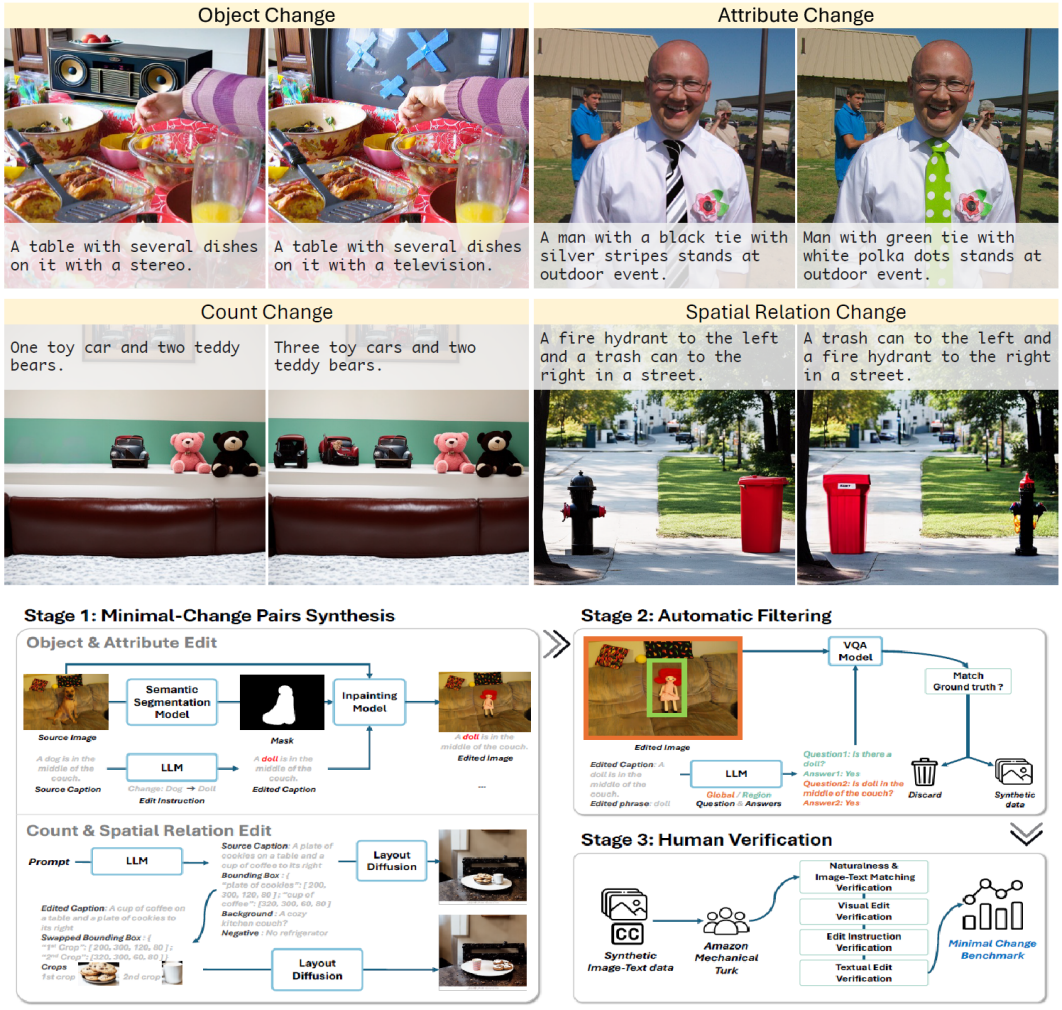


Figure 4: Overview of VisMin benchmark (top), and dataset creation pipeline (bottom). VisMin consists of four types of minimal-changes – object, attribute, count and spatial relation – between two image-captions pairs. The evaluation task requires a model to predict the correct image-caption match given: 1) two images and one caption, 2) captions and one image. Adapted from Awal et al. (2024).

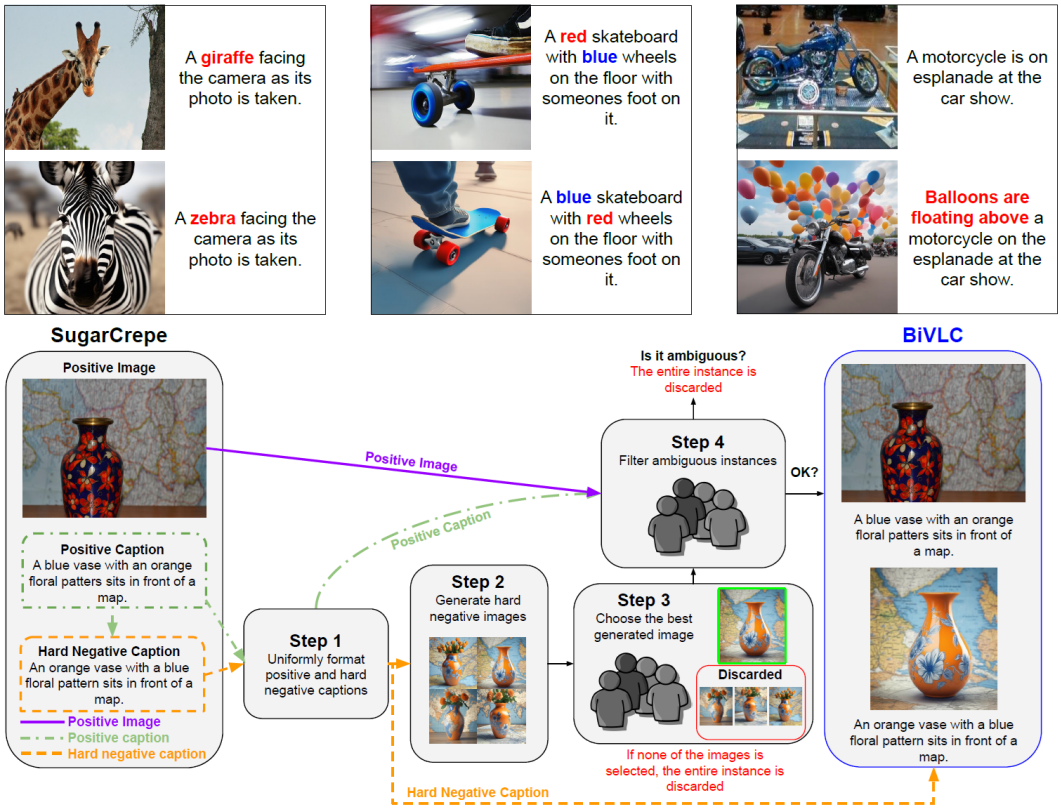


Figure 5: Overview of BiVLC benchmark (top), and dataset creation pipeline (bottom). Starting from SUGARCREPE instances, uniformly format positive and hard negative captions (Step 1), generate hard negative images (Step 2), ask human annotators to choose the best generated image (Step 3), and filter out ambiguous instances (Step 4). Adapted from Miranda et al. (2024).

4.1.5 Beyond Binary Retrieval

Even a benchmark with well-constructed negatives inherits the limits of the retrieval format itself. A choice between two candidates says little about which ability was missing when a model gets it wrong, and the score used to make that choice can be influenced by properties of the caption that have nothing to do with the image. The works in this group respond by changing the format rather than the data, either by fixing the scoring procedure so that it depends on visual evidence, or by replacing retrieval with tasks that ask the model something different.

SADE (Ma et al., 2024) addresses a problem specific to generative VLMs, where caption likelihood can be influenced by linguistic fluency rather than visual relevance. The benchmark introduces the **SyntaxBias Score** to measure this effect and proposes evaluation settings that reduce fluency advantages between positive and negative captions, including a content-only condition that tests whether models rely on semantic grounding rather than the language prior of the decoder.

$$\text{Score}_{\text{SyntaxBias}} = \Delta\left(\sum_{i=1}^{m} w_i \log P(p_i \mid p_{<i}; \theta) - \sum_{j=1}^{n} \hat{w}_j \log P(n_j \mid n_{<j}; \theta)\right).$$

The remaining benchmarks move further away from retrieval altogether. **ComPABench** (Li et al., 2026) evaluates whether models can compose skills learned separately by ensuring that individual capabilities are observed during training while their combinations are held out for testing. This design distinguishes genuine compositional generalization from memorization of previously seen combinations, a distinction that conventional benchmarks do not expose.

MMComposition (Hua et al., 2024) broadens evaluation in a different direction by covering single- and multi-image inputs, multiple question formats, and thirteen task categories (Figure 6). Rather than producing a single retrieval score, it provides a capability profile that reveals which aspects of compositional reasoning a model has mastered and which remain challenging.

4.1.6 New Compositional Axes

Most earlier benchmarks describe compositionality in terms of objects, their attributes, and the relations between them. That is one way concepts combine, but not the only one. The benchmarks in this group open further axes, meaning further kinds of structure that a model has to get right: spatial arrangement in three-dimensional scenes, the meaning of words that fuse into a single concept, and whether a description is matched to the correct part of the image rather than to the image as a whole.

SpaCE-10 (Gong et al., 2025) treats spatial understanding in realistic three-dimensional scenes as a compositional capability in its own right. The benchmark decomposes spatial competence into ten atomic skills—including counting, distance estimation, and multi-view reasoning—and constructs questions that require multiple skills simultaneously, allowing failures to be localized to individual capabilities and to their combinations.

Compun (Kumar et al., 2024) moves compositional evaluation below the sentence level by testing whether models interpret compound nouns such as "lab coat" as unified semantic concepts. Unlike most benchmarks in this survey, which examine failures to combine concepts that should remain separate, Compun targets the opposite error: treating a semantically fused expression as a simple combination of its individual words. **NegBench** (Alhamoud et al., 2025) opens a further axis in negation, testing whether models can represent what a caption states is *absent* — a construction that concept matching cannot handle in principle, since the negated object is mentioned

A complementary direction examines whether compositional understanding is reflected in visual grounding rather than only global similarity. **ARPGrounding** (Zeng et al., 2024) evaluates whether a model's visual focus corresponds to the region described by an attribute, relation, or priority phrase, revealing cases where correct retrieval is achieved using incorrect visual evidence.

$$\text{Mean Sample Activation} = \text{act}(M_i, H_j) = \frac{1}{\sum M_i} \sum M_i \odot H_j$$

$$\text{Sample Score} = f(M_0, H_0, M_1, H_1) = \begin{cases} 1, & \text{if } act(M_0, H_0) > act(M_1, H_0) \\ & \text{and } act(M_0, H_1) < act(M_1, H_1), \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Dataset Mean Accuracy} = \text{act} = \frac{1}{N} \sum f(M_0, H_0, M_1, H_1)$$

MultihopSpatial (Lee et al., 2026) extends this idea to multi-step spatial reasoning by requiring models to both solve a chain of spatial relations and ground the intermediate steps, distinguishing genuine compositional reasoning from shortcut-based answer selection.

4.1.7 Scoring Procedures

The last group in this section changes neither the images nor the captions. What it examines is the rule that turns a model's scores into a verdict, since the same predictions can be counted as a success or a failure depending on how strict that rule is. This makes it a benchmark contribution of a different kind: an existing evaluation is rerun unchanged, and only the definition of a correct answer is revised.

The most recent development in this literature leaves both the data and the model unchanged and instead revisits how predictions are converted into scores. Zhu et al. (2025) argue that Winoground's Group Score can underestimate model capability because its conjunctive requirement penalizes models that capture the

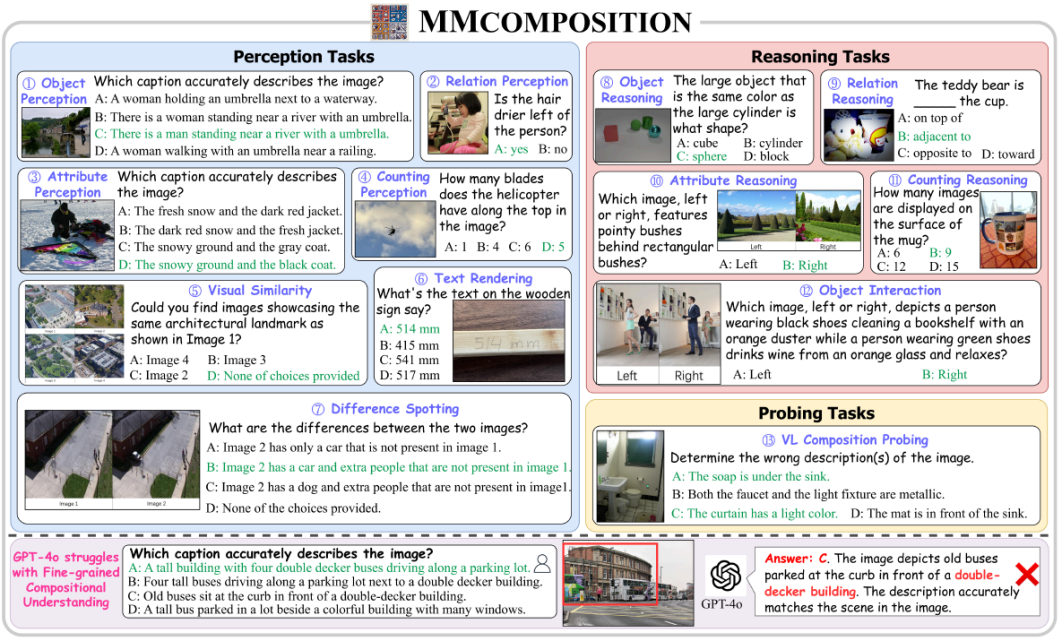


Figure 6: MMCOMPOSITION comprises 13 categories of high-quality VL composition QA pairs, covering a wide range of complex compositions. Adapted from Hua et al. (2024).

overall image–caption matching structure but fail one individual pairwise comparison. Their **GroupMatch** metric replaces this formulation with a global matching criterion that evaluates the best assignment between images and captions within a group, while **SimpleMatch** provides a more relaxed alternative. By changing only the evaluation rule, this work reveals that part of the reported compositional gap may originate from the metric itself rather than from the model representation.

$$\text{GroupScore}(s) = \begin{cases} 1, & \text{if } \forall i: \ s_{ii} > \max_{j \neq i} s_{ij} \text{ and } s_{ii} > \max_{j \neq i} s_{ji}, \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{GroupMatch}(s) = \begin{cases} 1, & \text{if } \sum_{i=1}^{k} s_{i,\pi^*(i)} > \sum_{i=1}^{k} s_{i,\pi(i)}, \quad \forall \pi \neq \pi^\star, \\ 0, & \text{otherwise.} \end{cases}$$

4.2 Analysis: Theoretical Accounts of Compositional Failure

Constructing benchmarks that reliably expose compositional failure is only half of the literature's contribution. The other half asks why the failures occur, and it is rarely the work of separate papers: a benchmark is typically accompanied by an investigation of the results it produces, so that the same study both establishes a deficit and offers an account of it. Analyses of this kind draw on two sources of evidence. Some are conducted on a benchmark the authors introduce themselves, where the construction of the data is designed from the outset to make a particular explanation testable; others are conducted on established benchmarks, using controlled perturbations, probes of internal representations, or comparisons across model families and training regimes to interrogate results the field already accepts.

What unites the twenty-five papers reviewed here is that each attempts to identify *where* the compositional deficit originates, and they disagree productively about the answer. The candidate locations span the whole pipeline: the objective a model is trained on, the component in which structural information is encoded, the data used to measure it, the rule used to score that measurement, and the geometry in which the resulting

representations are organized. These accounts are not mutually exclusive, and none has proved sufficient on its own; each explains part of the observed deficit and leaves a remainder for the next, so that the current understanding of the problem consists of the accumulated set rather than of any single explanation.

These analyses also carry directly into Section 4.3. An account of where a failure originates is, implicitly, a prescription for where an intervention should act, and most of the mitigations reviewed in this survey are motivated by one of the diagnoses presented here. The clusters that follow are therefore grouped by the location each account identifies. Table 2 summarizes this taxonomy.

4.2.1 Contrastive Learning as a Rational Shortcut

The first group of analyses looks at the objective the model is trained on. Its claim is that compositional weakness is not a defect that appeared by accident, but the expected result of training a model in a way that never requires structure. On this view the model is behaving correctly with respect to what it was asked to do, and the problem lies in what it was asked.

The starting point for subsequent analysis is Yuksekgonul et al. (2022)'s argument that standard contrastive pretraining provides little incentive to learn compositional structure. In large-scale image–text datasets, concept-level matching is often sufficient for retrieval, meaning that representing relationships between concepts contributes little additional signal to the contrastive objective. Bag-of-words behavior is therefore not an accidental deficiency but a rational outcome of optimization: models converge on representations that solve the training objective without requiring sensitivity to structure.

Tang et al. (2023) provide a more specific mechanism explaining how this shortcut appears during prediction. When a scene contains multiple objects with strongly associated attributes, a contrastive model attempts to account for the concepts present in the text. If one object's attributes are already explained by an explicitly mentioned object, remaining concepts may be incorrectly assigned to the queried object, producing attribute-binding errors. Their comparison across training objectives shows that this behavior is characteristic of contrastive learning rather than a general property of vision-language models, locating the source of the shortcut in the objective itself.

4.2.2 Architectural Bottlenecks

If the objective explains why structure is never learned, it does not say which part of the model fails to carry it. The analyses in this group open the model up and look inside, using probes that test what can still be recovered from a representation at a given stage. This turns a general claim about model behavior into a claim about a specific component, which matters because it determines where any repair has to be applied.

Kamath et al. (2023b) use reconstruction probes on the diagnostic framework described above, to show that key structural information—including spatial relations, temporal relations, and negation—is already missing from the text encoder representation before cross-modal comparison occurs. This finding places a fundamental constraint on mitigation: improving later fusion mechanisms cannot recover information that has already been discarded upstream.

Yun et al. (2023) examine a related but deeper question: whether models organize the primitive concepts required for composition in a meaningful way at all. Their **CompMap** framework probes VLM representations for attributes, part-level properties, and shapes, then evaluates both their usefulness for composite recognition and the interpretability of the resulting concept combinations. The analysis reveals a dissociation between these properties: VLM activations support strong downstream recognition, yet the inferred concept compositions often diverge from those derived from ground-truth concepts and rely on irrelevant primitives. The gap persists across model scales and prompt variations, suggesting that the limitation lies in how concepts are structured within pretrained representations rather than in the probing procedure itself.

4.2.3 Evaluation-Layer Problems

The analyses in this group argue that part of the measured failure does not belong to the model at all. Before concluding that a representation lacks structure, one has to rule out that something in the measurement

Category	Representative papers
Contrastive learning as a rational shortcut	Contrastive shortcut (Yuksekgonul et al., 2022) Concept association mechanism (Tang et al., 2023)
Architectural bottlenecks	Text-encoder bottleneck (Kamath et al., 2023b) Primitive-concept probing (Yun et al., 2023) BiVLC/TROHN (Miranda et al., 2024)
Evaluation-Layer problems	SugarCrepe (Hsieh et al., 2023) ConMe (Huang et al., 2024) SugarCrepe++ (Dumpala et al., 2024) Test-time matching (Zhu et al., 2025) NLI scoring (Cascante-Bonilla et al., 2024) SCRAMBLe (Mishra et al., 2025)
Mitigation side effects	Hard positives (Kamath et al., 2023a) Preserving multi-modal capabilities (Oh et al., 2024)
Training objectives and learning dynamics	ComPABench (Li et al., 2026) SADE (Ma et al., 2024) Iterated learning (Zheng et al., 2024) Causal graphical models (Parascandolo et al., 2025)
Recognition without structure	SPEC (Peng et al., 2024) CounterCurate (Zhang et al., 2024)
Capability-level bottlenecks	MMComposition (Hua et al., 2024) SpaCE-10 (Gong et al., 2025) Compun (Kumar et al., 2024) ARPGrounding (Zeng et al., 2024) MultihopSpatial (Lee et al., 2026)
Non-linear compositional geometry	Geodesic compositionality (Berasi et al., 2025)

Table 2: Summary of the analysis papers reviewed in Section 4.2.

produced the result: the data, the way incorrect candidates were written, the rule used to score the answer, or the wording through which the model was questioned. Each paper here takes one of these layers and shows what a model can appear to do, or fail to do, for reasons that have nothing to do with compositional ability.

Hsieh et al. (2023) open this line at the level of benchmark data. Rule-based perturbation introduces unintended statistical regularities—negatives that are physically implausible or linguistically unnatural—allowing a model without access to the image to identify the correct caption from text alone. The resulting gap between VLMs and text-only baselines therefore partly reflects benchmark construction rather than compositional capability. Huang et al. (2024) show that the problem persists after removing obvious linguistic artifacts: negatives generated by a language model can be fluent yet inconsistent with the actual image because the generator never observes the scene, allowing models to reject them through generic visual plausibility rather than compositional reasoning. Their error analysis across semantic categories further shows that failure patterns differ across architectures, suggesting that compositional weakness is distributed across multiple competencies, including attribute binding, counting, and spatial arrangement.

The critique then extends from negative construction to the evaluation format and scoring procedure. Dumpala et al. (2024) show that models passing binary discrimination may still lack sensitivity to semantic equivalence, with image-grounded scoring outperforming text-only evaluation on this stricter criterion. Improvements are concentrated in methods that explicitly strengthen the text representation, reinforcing the

view that part of the bottleneck lies in textual encoding rather than cross-modal fusion alone. Zhu et al. (2025) extend this argument to the metric itself, showing that near-random Group Scores can conflate two cases: genuinely lacking the required similarity structure, and possessing it in a form that the conjunctive metric does not credit. The difference between conjunctive and matching-based scoring reveals substantial hidden capability in both contrastive models and multimodal LLMs. Together with Hsieh et al. (2023), these findings suggest that reported compositional gaps contain at least three separable components: genuine representational failure, dataset artifacts, and metric-induced underestimation.

Cascante-Bonilla et al. (2024) identify a fourth problem in the textual scaffolding used to evaluate generative models. Approaches that ask an LLM to decompose captions into validation questions or scene graphs often preserve much of the original wording, allowing models that rely on lexical overlap to succeed without deeper compositional understanding. Moreover, the decomposing LLM never observes the image, meaning that any external world knowledge it introduces becomes an unverified assumption in the evaluation process. This failure mode is therefore orthogonal to the previous three: it concerns neither the representation, the data, nor the metric itself, but the intermediate textual interface through which the model is assessed.

4.2.4 Mitigation Side Effects

The analyses in this group look at what an intervention does besides what it was designed to do. Methods that reduce one failure can introduce another, either by teaching the model a habit that is too broad or by damaging abilities it already had. This belongs in the analysis section because it changes what a reported score means: once a model has been fine-tuned for compositionality, its behavior reflects that fine-tuning as much as its pretraining.

Kamath et al. (2023a) show that training a model to reject semantically altered captions does not produce a selective sensitivity to meaning. The same pressure suppresses similarity scores for captions that have been reworded while remaining semantically intact, so that surface variation and semantic change are conflated rather than distinguished. Their **Brittleness** metric makes the effect measurable and establishes that benchmark improvements obtained through hard-negative training can coexist with, and be caused by, degraded handling of valid paraphrases.

Oh et al. (2024) trace a second cost to its mechanism. Hard-negative fine-tuning consistently erodes zero-shot and retrieval performance, and they attribute this to the interaction between a global hard-negative loss and the single pooled representation that Kamath et al. (2023b) had already identified as a bottleneck: because a hard negative is encoded almost identically to its positive in one global vector, the gradient separating them necessarily perturbs the representation that downstream tasks depend on. The empirical signature is a systematic anti-correlation between compositional gains and multimodal task performance across prior methods. This contributes a fourth component to the decomposition above—alongside representational failure, dataset artifacts, and metric effects, the trade-off that mitigation itself imposes.

4.2.5 Training Objectives and Learning Dynamics

The accounts above take the training objective as given and ask what it fails to encourage. The analyses in this group treat it as something to be varied, comparing different ways of training the same model and observing which produce compositional behavior. What is varied differs from paper to paper: whether the model is tuned by supervision or by reward, how a candidate caption is scored, the order in which words are predicted, and whether training happens once or is repeated over successive rounds.

Li et al. (2026) isolate the choice between supervised fine-tuning and reinforcement learning as a determinant of whether independently mastered skills can be combined. Supervised fine-tuning produces near-perfect performance on individual subtasks while disrupting compositional generalization, whereas reinforcement learning preserves and often improves it—an asymmetry they attribute to supervised fine-tuning's tendency to memorize task-specific reasoning paths rather than transferable principles. The resolution is partial: reinforcement learning does not close the cross-modal gap in which skills learned on text fail to transfer to visually presented inputs, a residual failure they tie specifically to insufficient visual-to-text grounding during reasoning.

Ma et al. (2024) examine how the scoring objective shapes generative VLM behavior. Under controlled perturbation, likelihood-based scoring degrades sharply when fluency is disrupted yet remains largely stable when semantic content changes with grammatical form preserved, while contrastive scoring behaves in the opposite way; replacing real images with noise leaves generative accuracy nearly unchanged where contrastive accuracy collapses, indicating heavy reliance on the language model prior rather than on visual evidence. Because their fluency diagnostic can be applied to existing datasets, they further show the imbalance to be a property of the benchmark landscape rather than of any single model.

Two further analyses locate the lever outside the loss function. Zheng et al. (2024) import the cultural-transmission hypothesis from cognitive science, on which human compositionality emerges from the repeated pressure of teaching a language to new learners rather than from scale; standard VLM training, in which two encoders co-adapt continuously over a single lifetime, has no analogue of this pressure, which is consistent with the observation that larger models and datasets do not yield more compositional representations. Their supporting analysis shows that representations trained under an iterated procedure are measurably easier for a freshly initialized agent to acquire and that the model's estimated Lipschitz bound falls across generations, linking compositionality to a smoothness property of the representation rather than to any dataset- or objective-level artifact. Parascandolo et al. (2025) locate it in the prediction order of generative training, arguing that standard autoregressive factorization forces the absorption of spurious dependencies induced by English surface order—an adjective must be predicted before the noun it modifies is known—while fully parallel prediction discards inter-word dependencies altogether. Conditioning each word on its syntactic ancestors outperforms both extremes, indicating that the dependencies that matter are the sparse syntactic ones. A further ablation shows that discarding penultimate-layer visual features also costs compositional performance, a visual-side echo of the information-loss account of Kamath et al. (2023b).

4.2.6 Recognition Without Structure

A pattern recurs throughout this literature: models are reliable at identifying which things are present in a scene and unreliable at describing how those things are arranged. The analyses in this group show that the pattern is not only about language. The same split appears in properties that are purely visual, such as size, position, and quantity, where a model can name every object in an image and still fail to report how they are placed relative to one another.

Peng et al. (2024) document it for fine-grained visual concepts. Across contrastive and generative architectures, performance on size, position, count, and existence falls to near chance on images whose object categories the same models recognize reliably. The dissociation is diagnostic rather than merely descriptive: the models encode what is present and not how it is configured. They attribute this to the contrastive objective, which, because items in a randomly sampled batch differ in many respects at once, can be minimized by attending to nouns alone—reinforcing the shortcut account of Yuksekgonul et al. (2022) and Tang et al. (2023) and extending it from the textual to the visual side.

Zhang et al. (2024) reach a parallel conclusion for physically grounded reasoning. Contrastive and generative models alike perform at or barely above chance when distinguishing left/right and above/below arrangements or counting objects, including in cases where the model demonstrably identifies every object in the scene. Their explanation points to the supervision rather than the objective: web-scale captions rarely contrast scenes containing the same objects in different positions or quantities, so nothing in the training signal rewards representing those relations. Taken with Peng et al. (2024), this places the failure in the visual organization of scenes and not only in the structure of language.

4.2.7 Capability-Level Bottlenecks

Some benchmarks record which skills each question requires. This supports a different kind of analysis, in which a failure is attributed to a specific weak skill rather than to compositional ability in general. It also allows a second question to be asked: how much of the difficulty comes from the skills themselves, and how much from having to use several of them at once. The analyses in this group use that information to locate where compositional performance actually breaks down.

The broadest such analysis is that of Hua et al. (2024), which finds models reasonably competent at perceiving objects, attributes, and relations, but considerably weaker on counting, difference spotting, object interaction, and compositional probing—uniformly tasks requiring simultaneous reasoning over multiple visual elements. Their architectural ablations identify visual representation quality as the operative constraint: encoders that preserve resolution and aspect ratio improve compositional performance, whereas language decoder capacity yields diminishing returns beyond a threshold, past which the visual encoder becomes the dominant bottleneck. Gong et al. (2025) sharpen this into a capability-level result for three-dimensional scenes. Counting emerges as the constituent whose weakness most consistently degrades compositional performance, remaining far below every other spatial dimension and failing to improve reliably with parameter count. More telling is the compositional effect itself: adding capability requirements to a fixed question type produces accuracy drops far larger than the isolated difficulty of the added skills would predict, indicating that errors compound when skills must be chained. The pattern echoes the size, position, and count dissociation reported by Peng et al. (2024) and Zhang et al. (2024) in two dimensions and extends it to full 3D scenes.

Three analyses localize bottlenecks along the axes their benchmarks opened. Kumar et al. (2024) find compound-noun failures concentrated in attributed compounds, where one noun modifies another without itself being visually present, while compounds whose constituents are both visually absent are handled most reliably, apparently because they were encountered as holistic visual concepts during pretraining. The weakness is thus specific to cross-concept binding: familiar compounds stored as wholes can be retrieved, but interpretation fails when one noun must be understood as semantically transforming the other rather than co-occurring with it. Zeng et al. (2024) show that contrastive models often register the objects participating in a relation without encoding the relation, activating on both objects simultaneously when presented with a relational phrase; along the priority dimension, grounding follows visual salience and object frequency rather than grammatical role, a failure invisible to any retrieval metric that does not require localization. Lee et al. (2026) separate spatial failure into two components that standard benchmarks conflate—an inability to chain multiple spatial and attribute conditions, and an inability to localize the object corresponding to a selected answer—and find models that rank highly on conventional spatial benchmarks solving multi-hop queries through statistical shortcuts, as evidenced by near-zero grounding accuracy alongside reasonable answer selection. Their error analysis places the bottleneck not in spatial recognition but in the compositional chaining of inferences, with severity scaling as ego-centric perspective-taking is required.

4.2.8 Non-Linear Compositional Geometry

The last analysis in this section reverses the question the others ask. Rather than looking for the point at which compositional structure is lost, it asks whether the structure is present all along and simply not visible to the tools normally used to look for it. If that is the case, then part of the reported gap is a limitation of the analysis rather than of the model.

Berasi et al. (2025) argue that VLM visual embeddings already contain compositional structure, but that this structure is geodesic rather than linear. As a result, linearly decomposable approximations perform poorly, whereas geometry-aware decompositions that respect the underlying manifold can recover the relevant relationships. Their analysis identifies two challenges specific to visual representations—noise introduced by irrelevant background content and sparsity among certain concept combinations—and shows that addressing them through a weighted Riemannian decomposition enables recovery of compositional relationships for unseen attribute–object combinations. This perspective reframes the central question from whether VLMs contain compositional structure to where that structure resides, what geometric form it takes, and why standard evaluation procedures may fail to reveal it.

4.3 Mitigation: How Can Compositionality Be Improved

Beyond constructing benchmarks and diagnosing the sources of compositional failure, a substantial share of the literature proposes practical interventions. A *mitigation*, as the term is used here, is any deliberate modification intended to improve a model's compositional behavior, and the works reviewed in this section differ mainly in which part of the pipeline they modify. Some act on the training data, constructing examples

Category	Representative papers
Foundational hard-negative and cross-modal interventions	NegCLIP (Yuksekgonul et al., 2022) Cross-modal attention (Tang et al., 2023) Hard positives (Kamath et al., 2023a)
Efficient, capability-preserving fine-tuning	CLIC (Peleg et al., 2025) FSC-CLIP (Oh et al., 2024)
Visual hard negatives	VisMin (Awal et al., 2024) BiVLC/TROHN (Miranda et al., 2024) SPEC (Peng et al., 2024) CounterCurate (Zhang et al., 2024) SPARCL (Li & Li, 2025) SCRAMBLe (Mishra et al., 2025)
Redesigning the training objective	RL-Ground (Li et al., 2026) Iterated learning (Zheng et al., 2024) Causal graphical models / COGT (Parascandolo et al., 2025)
Axis-specific interventions	ARPGrounding (Zeng et al., 2024) Retrieval with example captions (Kumar et al., 2024) SADE bias mitigation (Ma et al., 2024)
Diagnostic-derived architectural and data recommendations	MMComposition (Hua et al., 2024) SpaCE-10 (Gong et al., 2025) NegBench (Alhamoud et al., 2025)
Inference-time methods	Test-time matching (Zhu et al., 2025) CECE (Cascante-Bonilla et al., 2024) COCO-Tree (Sinha et al., 2025) GDE (Berasi et al., 2025)

Table 3: Summary of the mitigation papers reviewed in Section 4.3.

that a model cannot resolve by recognizing concepts alone. Some act on the loss or the training objective, changing what the model is rewarded for rather than what it is shown. Some act on the architecture, adding components or altering how information flows between the modalities. Some target a single diagnosed failure rather than compositional ability in general. And some leave the model entirely untouched, intervening instead on how its existing representations are queried at inference time.

These are best read as several angles on the same problem rather than as a single line of descent. A given intervention is usually motivated by one of the diagnoses of Section 4.2, and where a diagnosis locates the failure largely determines where the corresponding method acts, so approaches developed independently can converge on similar mechanisms while others proceed on entirely separate assumptions.

A constraint recurs across these approaches regardless of where they intervene. The models being modified are already strong at the tasks they were trained for, and compositional gains are of limited value if they are purchased by degrading zero-shot classification or retrieval. Several of the methods reviewed here are designed explicitly around this constraint, restricting which parameters are updated, reformulating the loss to avoid disturbing the pooled representation, or avoiding parameter updates altogether. Table 3 summarizes this taxonomy.

4.3.1 Foundational Hard-negative and Cross-modal Interventions

During contrastive training, the incorrect captions a model is compared against are the other captions in the batch, and these usually describe completely different scenes. A model can therefore rank the correct caption first simply by checking which objects are mentioned, without ever using word order or relations. Since that already works, there is no pressure to learn anything more, and this is the shortcut that the analyses in Section 4.2 identify. The earliest mitigations remove it directly, by placing captions in the batch that mention the same things in a different arrangement, so that concept matching no longer separates the right answer from the wrong one.

NegCLIP (Yuksekgonul et al., 2022) implements this by adding two kinds of hard negative to each batch, captions whose structure has been altered while their words are preserved, and visually similar images retrieved as nearest neighbors in embedding space. Penalizing solutions available through concept matching alone improves relational understanding and word-order sensitivity while largely preserving downstream classification and retrieval, establishing hard-negative supervision as the default starting point for following works; the method was evaluated as a fine-tuning procedure rather than at pretraining scale.

Two recent interventions explore orthogonal modifications to the training recipe. Rather than modifying the data, Tang et al. (2023) modify the architecture, adding a cross-modal attention module on top of the CLIP encoders and fine-tuning it on visual question answering, on the reasoning that concept association bias reflects insufficient interaction between the modalities; the reduction in measured bias and the improvement in VQA performance track each other across model variants, tying compositional gains to stronger visual grounding. Kamath et al. (2023a) instead address the oversensitivity that hard-negative training induces, augmenting each batch with hard positives—paraphrases and synonym substitutions that preserve meaning and should therefore retain high similarity—alongside the negatives. Applied to a large set of training pairs derived from COCO covering synonym replacement and valid attribute reassignment, the joint objective exposes the model to the contrast between surface change and semantic change rather than to negative pressure alone, yielding a more selective representation than hard negatives can produce by themselves.

4.3.2 Efficient, Capability-preserving Fine-tuning

A common limitation is evident across the foundational interventions: compositional gains come with weaker zero-shot classification and retrieval, and often do not transfer to benchmarks that require understanding meaning rather than spotting a changed word. The methods in this group treat that cost as part of what they are designed to solve rather than as an acceptable side effect. What distinguishes them is therefore not only how they improve compositionality, but what they do to avoid disturbing the abilities the model already has.

CLIC (Peleg et al., 2025) restructures the supervision itself. Each training step concatenates two images side by side and derives supervision from their dense captions, forming several semantically equivalent positives by sampling sentence combinations across the two caption sets and a hard negative by exchanging a single word between sentences within the same part-of-speech category, using a lightweight tagger and requiring neither a language model nor a diffusion model. Three losses are combined—a contrastive loss over the positives, a hard-negative loss on each positive–negative pair, and a unimodal text loss penalizing sensitivity to caption order—and only the text encoder is updated, with the vision encoder frozen and training alternating between the concatenated-image regime and standard single-image CLIP training to prevent forgetting. **FSC-CLIP** (Oh et al., 2024) attacks the same problem at the point their analysis identifies as its source. Since contrasting a hard negative against its positive in a single pooled vector is what distorts the multimodal representation, the global hard-negative loss is replaced with a local one computed over token–patch alignments, which exposes the difference without collapsing the global representation, and supplemented with selective calibrated regularization, a focal weighting that concentrates supervision on genuinely confusing negatives together with label smoothing that grants hard negatives a small positive margin rather than treating them as strictly negative. Both methods reframe the goal of compositional fine-tuning from maximizing a benchmark score to preserving the model's full multimodal competence.

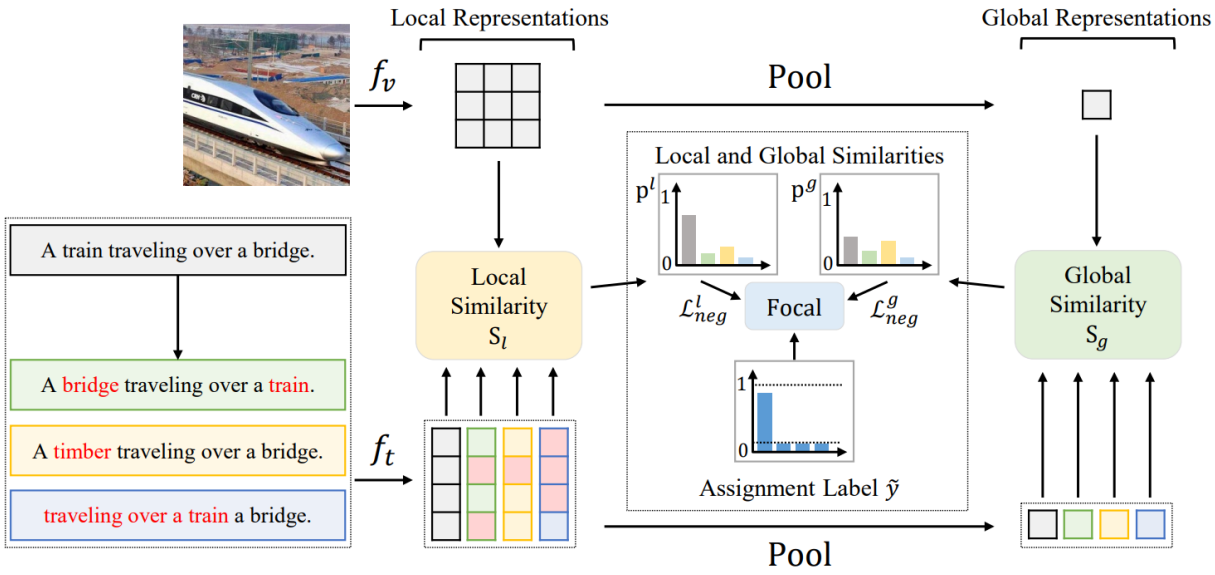


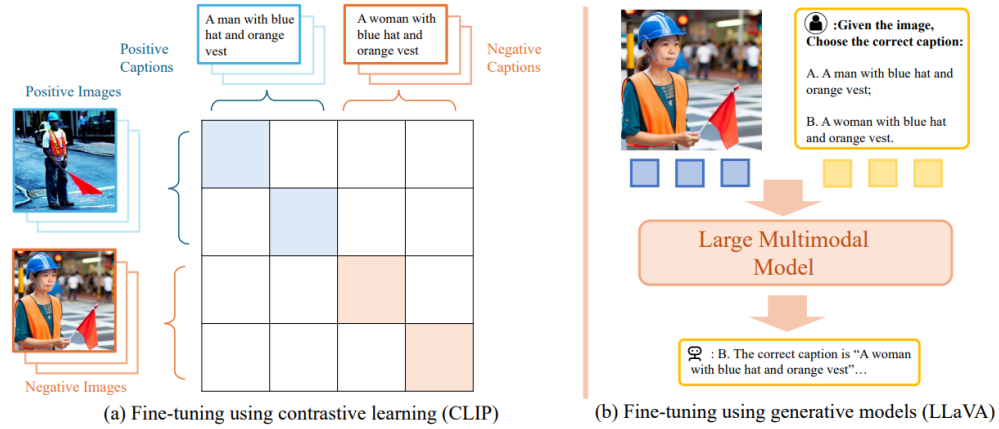
Figure 7: FSC-CLIP framework that incorporates Local Hard Negative (LHN) Loss with Selective Calibrated Regularization (SCR), alongside a global HN loss. The LHN loss measures similarity between an image and a text at the patch and token levels to more accurately identify subtle differences between original and HN texts. SCR combines focal loss with label smoothing to mitigate the adverse effects of using hard negative losses. Adapted from Oh et al. (2024).

4.3.3 Visual Hard Negatives

Training against altered captions teaches a model to be careful about text, and evaluation shows that this does not carry over to finding the right image for a description (Section 4.1). The methods in this group respond by making the negatives visual, training the model against images that differ from the correct one in a single compositional respect. Most of the work in this line concerns how to produce such images at all, since an edit has to change the intended property while leaving everything else in the picture intact.

Awal et al. (2024) take the first step by repurposing the VisMin construction pipeline as a training data generator, producing tens of thousands of minimal-change image–text pairs and using them to fine-tune both CLIP and a multimodal LLM with no change to architecture or loss. Miranda et al. (2024) make the argument explicit, contending that hard negative texts alone cannot address the text-to-image weakness, and construct two COCO-derived training sets to demonstrate the point: one supplying hard negative captions and one in which the best of those captions, filtered for plausibility and linguistic acceptability, are rendered into images. Peng et al. (2024) arrive at the same conclusion from the optimization side rather than the data side, introducing a hard-negative-aware contrastive loss that treats the confusing images and texts within a candidate set as negatives in both the image-to-text and text-to-image terms, so that the vision and text encoders are pushed to discriminate subtle differences simultaneously, and combining it with a standard CLIP loss on web data to protect zero-shot ability.

Later work in this line is preoccupied less with the principle than with the difficulty of generating usable images. **CounterCurate** (Zhang et al., 2024) extends the pipeline to physically grounded reasoning by matching the edit to the phenomenon: positional negatives are produced by flipping a directional keyword and mirroring the image, above/below and counting negatives by editing bounding boxes and re-rendering with a grounded generator, and attribute negatives by prompting GPT-4V for hard negative captions that a text-to-image model then renders, Figure 8(top). A grouping strategy forcing each positive to be distinguished from its own corresponding negative image and caption proves essential in ablation, and both contrastive and generative models are fine-tuned within the same pipeline. **SPARCL** (Li & Li, 2025), Figure 8(bottom), confronts the remaining obstacles of efficiency, alignment at the site of the intended change, and fidelity



(a) Fine-tuning using contrastive learning (CLIP) (b) Fine-tuning using generative models (LLaVA)

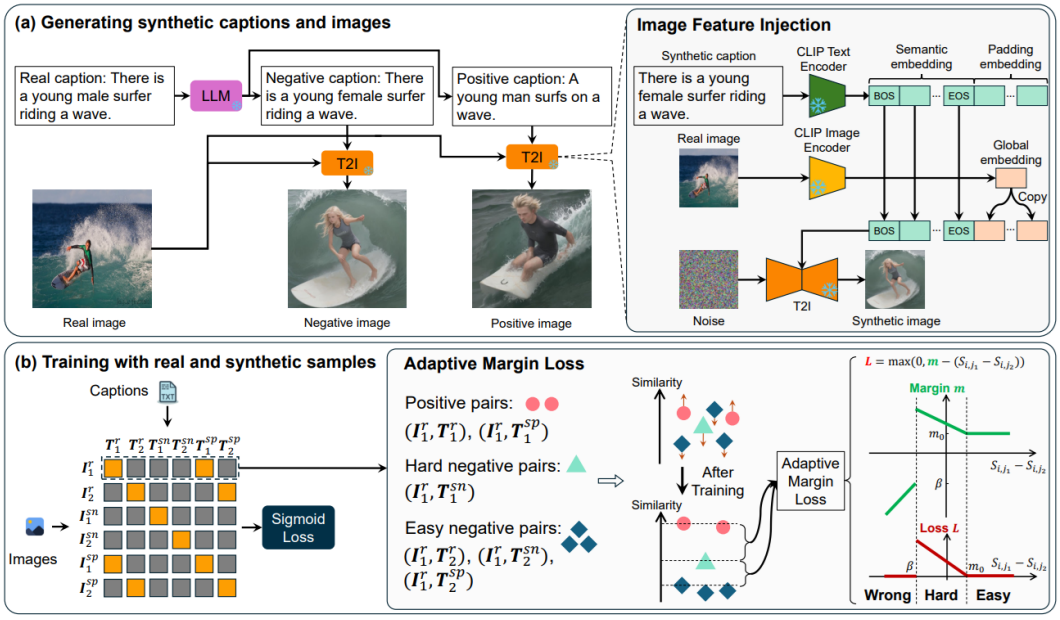


Figure 8: An overview of visual hard negatives frameworks. CounterCurate (top). SPARCL (bottom). Adapted from (Zhang et al., 2024) and (Li & Li, 2025), respectively.

elsewhere. Starting from real COCO pairs, it prompts an LLM for a hard negative and a meaning-preserving hard positive caption and generates images with a fast few-step model, keeps edits local by injecting the real image's features into the padding embeddings of the text prompt and closing the residual domain gap with a style-transfer step, and handles variable generation quality through an adaptive margin loss that separates positive, hard-negative, and easy-negative pairs and filters likely-incorrect generations by collapsing the margin when a purported positive is less similar than its negative. Notably, its ablations find that uncontrolled synthetic variation misleads compositional learning much as rule-based perturbation does, which is the artifact critique of Hsieh et al. (2023) restated on the training side. **SCRAMBLe** (Mishra et al., 2025) carries the same concern into the generative setting, balancing generated negatives for plausibility and grammaticality before preference tuning a multimodal LLM on them.

4.3.4 Redesigning the Training Objective

The methods above all work by supplying harder examples. The methods in this group change something else instead: what the model is asked to predict, how it is rewarded for predicting it, or how training is

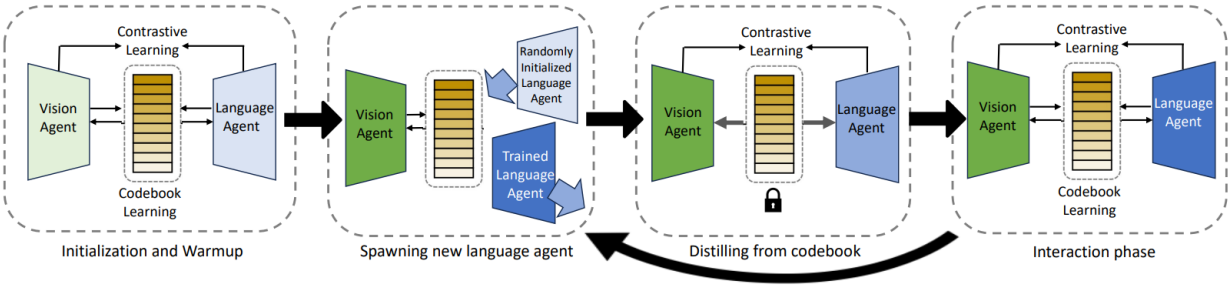


Figure 9: Iterated learning algorithm from IL-CLIP, is built on CLIP augmented with a shared codebook. Adapted from (Zheng et al., 2024).

organized over time. The aim is to make compositional structure something the model is trained on directly, rather than something it might pick up while learning to tell examples apart.

RL-Ground (Li et al., 2026) follows from the observation that reward design determines whether independently acquired skills can be combined. It modifies both the input format and the learning signal: a caption-before-thinking format requires the model to first verbalize the visual scene before entering its reasoning chain, introducing an intermediate textual representation that reduces the difficulty of multimodal reasoning, while a progress reward provides dense supervision over intermediate subgoal computations rather than rewarding only the final answer. Ablation shows that the two components are complementary, and that reinforcement learning initialized from a supervised checkpoint does not recover compositional ability because the task-specific biases introduced during supervised fine-tuning constrain the exploration of more compositional strategies.

Supervised Fine-Tuning:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{t=1}^{T} \log p_\theta \left(y_t \mid x, y_{<t}\right)$$

Reinforcement Learning with GRPO:

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^{(i)}|}\sum_{t=1}^{|o^{(i)}|}\left[\frac{\pi_\theta\left(o_t^{(i)} \mid q, o_{<t}^{(i)}\right)}{\pi_\theta\left(o_t^{(i)} \mid q, o_{<t}^{(i)}\right)} \cdot A(i,t) - \beta \cdot \mathrm{KL}\left(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\right)\right]$$

IL-CLIP (Zheng et al., 2024) intervenes on learning dynamics rather than the training examples themselves. Contrastive learning is reframed as a signaling game between a vision agent and a language agent communicating through a shared learnable codebook, Figure 9; the language agent is periodically discarded, reinitialized, briefly distilled from the frozen codebook, and then jointly trained again. Each cycle simulates a new generation that must relearn the communication protocol, encouraging representations that remain easy to acquire and, under the cultural-transmission hypothesis, more compositional. Ablations show that each part of the procedure contributes to the effect: resetting the vision agent instead, unfreezing the codebook during distillation, or removing the codebook entirely all reduce the resulting improvement.

COGT (Parascandolo et al., 2025) replaces contrastive scoring with a generative objective whose prediction structure follows syntax rather than surface word order. A dependency tree, Figure10(left), is extracted from each caption and used as a structured factorization, where each word is generated from its syntactic ancestors, syntactic type, and visual features rather than from all previous tokens. Implemented as a lightweight dependency-guided decoder, Figure10(right), on top of a frozen CLIP visual encoder, the method scores candidate captions through partially ordered likelihoods and transfers across visual backbones without architectural changes. The result suggests that injecting explicit linguistic structure into the prediction process can provide a complementary path toward compositionality beyond increasingly difficult hard-negative mining.

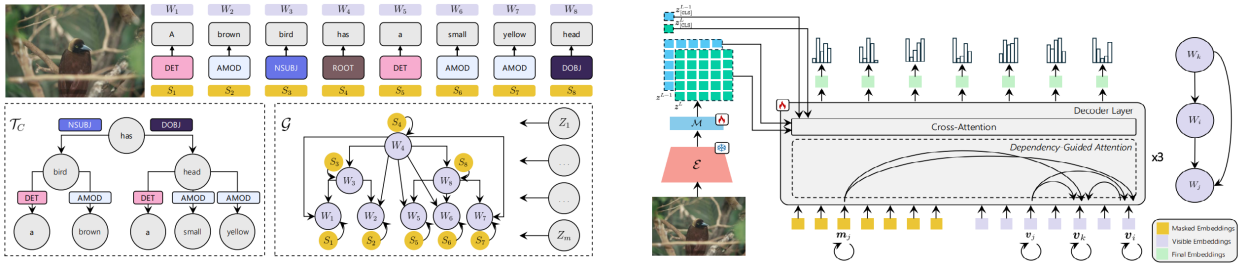


Figure 10: Dependency tree created by COGT, representing relations between words in a sentence (left). A schematic illustration of COGT's decoder (right). Adapted from (Parascandolo et al., 2025).

4.3.5 Axis-Specific Interventions

The previous methods try to improve compositional ability in general. The papers in this group do the opposite, taking a single failure that an earlier analysis identified and addressing that failure directly. The trade-off is deliberate: they give up broad coverage in exchange for acting on a known cause, which also makes them comparatively cheap, since several require no retraining at all.

For spatial grounding, Zeng et al. (2024) propose a composition-aware fine-tuning pipeline that requires no dense bounding box annotation. Dependency parsing is used to extract, from each training image's existing captions, text pairs describing distinct objects, and a pretext loss penalizes overlap between the Grad-CAM heatmaps the two texts produce, encouraging spatially distinct activations for descriptions differing in attribute, relation, or priority. That standard contrastive or image–text matching fine-tuning on the same data does not produce the same gains indicates the heatmap-diversity objective rather than the additional data is doing the work. For compound nouns, Kumar et al. (2024) propose **Retrieval with Example Captions**, a training-free alternative to minimal template prompts: several diverse LLM-generated captions place the compound in different situational contexts and the image with the highest mean similarity across them is selected, on the hypothesis that contextually varied keywords raise the activation of the compound-depicting image above its constituent-noun distractors. The gains concentrate in exactly the attributed-compound category their analysis identifies as the weakest, showing that prompt-level enrichment can partially compensate for a binding failure internal to the model.

Ma et al. (2024) intervene on measurement rather than on the model, on the grounds that a benchmark rewarding fluency cannot be used to certify grounding. Score-guided filtering removes candidate pairs in which the positive caption enjoys a statistically significant fluency advantage, retaining only those where the image must drive the decision, while a content-only challenge strips positives of syntactic scaffolding and pairs them against fluent but semantically unrelated negatives, deliberately inverting the usual advantage so that a model relying on the language model prior is penalized rather than rewarded. Human raters confirm the resulting benchmark achieves near-zero fluency disparity.

4.3.6 Diagnostic-derived Architectural and Data Recommendations

Not every contribution to this literature is a method. The studies in this group ran large diagnostic evaluations and turned the results into recommendations about how future models and training sets should be built. They are recorded separately because their advice concerns decisions that the training-level methods above simply inherit, such as how much visual detail an encoder preserves and what a fine-tuning set should contain.

Hua et al. (2024) argue from their architectural ablations that visual encoder design should prioritize preserving image resolution and aspect ratio, since models that downsample aggressively fall behind as task complexity increases; that scaling the language decoder improves compositional reasoning only up to a threshold beyond which the visual encoder becomes the binding constraint, so effort is better redirected toward higher-fidelity visual representations than toward continued LLM scaling; and that increasing the volume and compositional diversity of fine-tuning data yields consistent gains, indicating that current corpora

under-represent the multi-element scenarios the benchmark exposes. Gong et al. (2025) derive a narrower and sharper prescription from their capability-level diagnosis. Because counting is the constituent whose weakness most consistently degrades compositional performance, they advocate counting-focused supervision built from scenes that deliberately stress occlusion, crowding, and multi-view consistency rather than the clean object-count data typical of existing corpora; and because the observed collapse occurs specifically when view-conditioned relational chains must be executed together, they argue for capability-composition aware post-training that targets spatial relations, counting, and situated observation jointly rather than each in isolation. Both recommendations follow from the finding that scale alone does not close the gap, and converge with the objective-level conclusion of Li et al. (2026) that compositional ability must be targeted explicitly during training rather than expected to emerge from mastery of the individual subtasks.

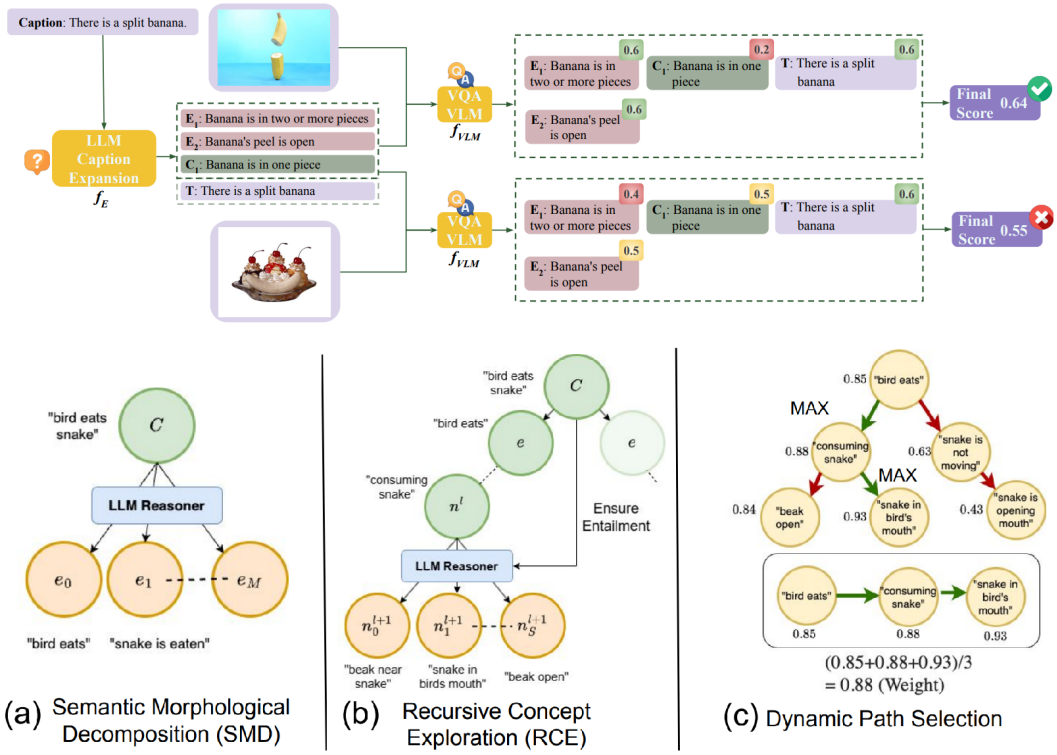


Figure 11: An overview of inference-time methods frameworks. CECE (top). COCO-Tree (bottom). Adapted from (Cascante-Bonilla et al., 2024) and (Sinha et al., 2025), respectively.

4.3.7 Inference-Time Methods

The most recent group of methods starts from a suggestion made by the analyses above: that the structure a model needs may already be present in its representations and simply hard to get at. These methods change nothing about the model. They work on the way it is used at test time, altering how a caption is presented, how its scores are combined, or how its embeddings are broken down.

Test-Time Matching (Zhu et al., 2025) converts their metric analysis into a self-improvement procedure. Each test group's highest-scoring image–caption matching is computed, matchings whose margin over the runner-up exceeds a confidence threshold are retained as pseudo-labels, the model is fine-tuned on them, and the process repeats with a decaying threshold so that coverage expands as the model improves; a global variant replaces the group structure with a single assignment problem solved with the Hungarian algorithm, extending the approach to non-grouped settings. That the gains exceed what the metric correction alone explains, and persist on benchmarks where conjunctive and matching-based scoring coincide, indicates that a meaningful fraction of the compositional gap is recoverable by better eliciting what pretrained models already represent.

Two methods intervene on the textual side of inference instead. **CECE** (Cascante-Bonilla et al., 2024), Figure11(top), replaces question-style decomposition with the inference relations of natural language inference: treating the caption as a premise, an LLM generates entailments, whose truth follows from the premise, and contradictions, which cannot hold if the premise does—with negation words explicitly disallowed, so that a contradiction must be expressed through different content rather than a syntactic marker. Because entailments and contradictions restate the scene rather than interrogate its tokens, the expansions are lexically far from the original, which is what breaks the keyword-overlap shortcut that decomposition inherits; scoring aggregates the probability assigned to affirming each entailment and denying each contradiction alongside the original caption, and their ablation shows the two cue types to carry complementary rather than redundant signal. **COCO-Tree** (Sinha et al., 2025) augments inference with an external reasoner in a neurosymbolic framework, Figure11(bottom), first decomposing a caption into disentangled morphological entities, then recursively expanding each into a hierarchy of semantically related concepts using an LLM, and finally applying a beam-search-inspired path-finding algorithm to identify the reasoning pathway through the concept tree that best combines visual grounding with linguistic entailment scores, blending this deliberative output with the VLM's original prediction. Beyond its accuracy gains, the selected path constitutes an interpretable rationale, an advantage over LLM-augmentation approaches that improve scores without exposing why.

Geodesically Decomposable Embeddings (Berasi et al., 2025) closes this line by acting on the representation geometry their analysis identified. Rather than assuming compositional concepts correspond to linear directions, GDE computes an optimal set of primitive direction vectors in the tangent space at the intrinsic mean of the embedding distribution, handling visual noise through a weighted averaging scheme based on CLIP image-to-text probabilities and sparsity by averaging over all available examples sharing each primitive, which allows classifiers for unseen attribute–object combinations to be constructed from seen ones without any training. That the same decomposition also yields debiased representations for group robustness supports the broader claim of this cluster: improving compositionality is sometimes a matter of gaining better access to representations the model already has, rather than of modifying the model further.

5 Comparison and Discussion

Section 4 took the papers one at a time. This section puts them side by side. The purpose is to ask questions that no individual paper answers: whether the benchmarks measure the same thing, whether the analyses agree on what is wrong, whether the mitigations can be compared to one another at all, and what happens when the answers to these questions are read together.

The comparison proceeds in four steps. It first maps the papers onto the three categories of Section 4, since how they overlap says something about how the field works. It then compares the benchmarks, the analyses, and the mitigations in turn, each against the others in its own category. It closes with the tensions that emerge from this exercise, which are the starting point for Section 6.

5.1 Mapping the Papers

Table 4 records which of the three categories each reviewed paper contributes to. Papers are grouped by the pattern of contribution rather than by date, so that the shape of the field is visible at a glance.

Twelve of the thirty-one papers contribute to all three categories. These papers introduce a benchmark, use it to work out what is going wrong, and propose a method to fix it. Only five papers contribute to a single category. The typical contribution in this field is therefore not a benchmark, an analysis, or a method on its own, but all three together.

This has a consequence worth stating plainly. When the same paper builds the test, explains the result, and supplies the fix, the fix is evaluated on a test designed by the people who already knew what they were going to build. Nothing improper follows from this, but the evidence is weaker than three separate results would be, because the assumptions behind the benchmark and the assumptions behind the method are the same assumptions. The papers in a single category are useful precisely because they break this pattern: Yun et al. (2023) analyzes without proposing a fix, while Peleg et al. (2025), Li & Li (2025), and Sinha et al. (2025) propose methods tested on benchmarks they did not design.

Paper	Named artifact / method	Contribution		
		B	**A**	**M**
All three categories				
Yuksekgonul et al. (2022)	ARO; NegCLIP	✓	✓	✓
Tang et al. (2023)	NCD/UNCD; cross-modal attention	✓	✓	✓
Kamath et al. (2023a)	ATA; brittleness; hard positives	✓	✓	✓
Miranda et al. (2024)	BiVLC; TROHN-TEXT/TROHN-IMG	✓	✓	✓
Peng et al. (2024)	SPEC; bidirectional negative loss	✓	✓	✓
Ma et al. (2024)	SADE; SyntaxBias Score	✓	✓	✓
Hua et al. (2024)	MMComposition	✓	✓	✓
Kumar et al. (2024)	Compun; retrieval with example captions	✓	✓	✓
Zeng et al. (2024)	ARPGrounding; heatmap pretext	✓	✓	✓
Gong et al. (2025)	SpaCE-10	✓	✓	✓
Zhu et al. (2025)	GroupMatch/SimpleMatch; TTM	✓	✓	✓
Li et al. (2026)	ComPABench; RL-Ground	✓	✓	✓
Benchmark and analysis				
Kamath et al. (2023b)	CompPrompts; ControlledImCaps	✓	✓	
Hsieh et al. (2023)	SugarCrepe	✓	✓	
Dumpala et al. (2024)	SugarCrepe++	✓	✓	
Huang et al. (2024)	ConMe	✓	✓	
Lee et al. (2026)	MultihopSpatial	✓	✓	
Analysis and mitigation				
Zhang et al. (2024)	CounterCurate		✓	✓
Oh et al. (2024)	FSC-CLIP		✓	✓
Zheng et al. (2024)	IL-CLIP		✓	✓
Cascante-Bonilla et al. (2024)	CECE		✓	✓
Berasi et al. (2025)	GDE		✓	✓
Parascandolo et al. (2025)	COGT		✓	✓
Mishra et al. (2025)	SCRAMBLe		✓	✓
Benchmark and mitigation				
Awal et al. (2024)	VisMin (benchmark + training data)	✓		✓
Alhamoud et al. (2025)	NegBench	✓		✓
Single category				
Thrush et al. (2022)	Winoground	✓		
Yun et al. (2023)	CompMap		✓	
Peleg et al. (2025)	CLIC			✓
Li & Li (2025)	SPARCL			✓
Sinha et al. (2025)	COCO-Tree			✓
Total (31 papers)		20	25	24

Table 4: Category membership of the reviewed papers, grouped by contribution pattern. B = benchmark or evaluation (Section 4.1), A = analysis of compositional failure (Section 4.2), M = proposed mitigation (Section 4.3). Twelve papers contribute to all three categories; five contribute to one.

5.2 Benchmark Comparison

Benchmarks are routinely cited together as evidence for the same claim, but they differ in ways that change the number each one reports. Table 5 sets out five of these differences: how large the benchmark is, which

Benchmark	Scale	Axes covered	Direction	Negative source	Artifact ctrl.
Foundational probes					
Winoground (Thrush et al., 2022)	400	word order	I2T/T2I	hand-curated	(✓)
ARO (Yuksekgonul et al., 2022)	50K+	relation attribute order	I2T	rule-based	×
Artifact correction					
SugarCrepe (Hsieh et al., 2023)	7.5K	replace swap add	I2T	LLM-generated	✓
SugarCrepe++ (Dumpala et al., 2024)	4.8K	replace swap add paraphrase equiv.	I2T text-only	LLM-generated	✓
ConMe (Huang et al., 2024)	24K	attribute object relation	I2T	VLM-panel dialogue	✓
Visual and bidirectional					
VisMin (Awal et al., 2024)	–	object attribute count spatial	I2T/T2I	LLM + diffusion	✓
BiVLC (Miranda et al., 2024)	2.9K	replace swap add	I2T/T2I	diffusion + human	✓
SPEC (Peng et al., 2024)	3K	size position existence count	I2T/T2I	synthetic engine	✓
Beyond retrieval and new axes					
MMComposition (Hua et al., 2024)	4.3K	13 categories	MCQ (single/multi)	human-annotated	✓
NegBench (Alhamoud et al., 2025)	79K	negation	retrieval MCQ	LLM + detector	✓
SpaCE-10 (Gong et al., 2025)	5K+	10 atomic 8 composite	MCQ (2D/3D)	human pipeline	✓

Table 5: Construction properties of the major compositional benchmarks, grouped by cluster. "Direction" indicates whether the benchmark supports image-to-text (I2T), text-to-image (T2I), or non-retrieval multiple-choice (MCQ) evaluation. "Artifact ctrl." reflects whether construction explicitly filters for the shortcuts identified by Hsieh et al. (2023): ✓ yes, (✓) partially, × no. "–" denotes a scale not reported in the source paper.

compositional properties it covers, which retrieval direction it exercises, how its incorrect candidates were produced, and whether its construction filters for the shortcuts described in Section 4.1. Rows are grouped by the cluster each benchmark belongs to, which shows the direction of travel: negatives move from rules to language models to image-conditioned generation, and coverage moves from captions alone to images and to formats other than retrieval.

A sixth difference is easy to miss because it is rarely presented as a design choice. Every benchmark also decides what counts as a correct answer, and these decisions do not agree. Table 6 groups the reviewed

Benchmark	Reported metric
Pairwise and n-way matching: rank the correct candidate above the incorrect ones	
ARO (Yuksekgonul et al., 2022)	Accuracy per subset
SugarCrepe (Hsieh et al., 2023)	Accuracy per replace/swap/add
VisMin (Awal et al., 2024)	Accuracy per change type, both directions
SPEC (Peng et al., 2024)	I2T and T2I accuracy (n-way)
Compun (Kumar et al., 2024)	Retrieval accuracy
Conjunctive group: several comparisons must succeed together	
Winoground (Thrush et al., 2022)	Text, Image, Group score
BiVLC (Miranda et al., 2024)	Text, Image, Group score (bidirectional)
Global matching: credit the best assignment over the candidate set	
Test-time matching / GroupMatch (Zhu et al., 2025)	Matching accuracy over candidate set
Two-sided: accept a preserved meaning and reject an altered one	
SugarCrepe++ (Dumpala et al., 2024)	Triplet (TOT) accuracy
Hard positives (Kamath et al., 2023a)	ATA; brittleness
Multiple choice: select among n options, reported per capability	
ConMe (Huang et al., 2024)	Question accuracy
MMComposition (Hua et al., 2024)	Accuracy over 13 categories
SpaCE-10 (Gong et al., 2025)	Accuracy per atomic and composite capability
ComPABench (Li et al., 2026)	In-distribution and OOD composition accuracy
Likelihood-based: score the caption by the probability of its words	
SADE (Ma et al., 2024)	SyntaxBias Score; subtask and content-only accuracy
Localization: check where the model looked, not how it ranked	
ARPGrounding (Zeng et al., 2024)	Grounding accuracy (attribute, relation, priority)
MultihopSpatial (Lee et al., 2026)	Answer accuracy; Acc@50 IoU
Paired-condition diagnostics: compare accuracy across matched settings	
NCD/UNCD (Tang et al., 2023)	Concept-association-bias score
CompPrompts/ControlledImCaps (Kamath et al., 2023b)	Text-only vs. multimodal accuracy; reconstruction probes

Table 6: Scoring rules across the reviewed benchmarks, grouped by kind. Eight kinds appear, and no benchmark reports its results under more than one of them.

benchmarks by the kind of scoring rule they use. Eight kinds appear. Some ask a model to rank one caption above another. Some require several such comparisons to succeed at once. Some ask for the best overall assignment between images and captions. Some require the model to accept one alternative and reject another in the same decision. Some score a caption by the probability the model assigns to its words. Some ignore ranking entirely and ask where in the image the model looked.

Each of these is a reasonable definition of compositional success, and each measures something a little different. The problem is that no benchmark reviewed here reports its results under more than one of them. A model that looks weak under a strict conjunctive rule may look competent under a matching rule applied to the same predictions, and nothing in the literature makes that difference visible, because the comparison is never run. Accuracy figures quoted across papers are therefore not directly comparable, even when the underlying data is similar.

	Source of failure						
Paper	**Obj.**	**Text**	**Vis.**	**Data**	**Eval.**	**Cap.**	**Geom.**
Yuksekgonul et al. (2022)	✓						
Tang et al. (2023)	✓						
Kamath et al. (2023a)	✓						
Peng et al. (2024)	✓		✓				
Ma et al. (2024)	✓				✓		
Oh et al. (2024)	✓	✓					
Zheng et al. (2024)	✓						
Parascandolo et al. (2025)	✓		✓				
Li et al. (2026)	✓		✓				
Kamath et al. (2023b)		✓					
Dumpala et al. (2024)		✓			✓		
Yun et al. (2023)			✓				
Miranda et al. (2024)			✓				
Hua et al. (2024)			✓	✓			
Zeng et al. (2024)			✓			✓	
Lee et al. (2026)			✓			✓	
Zhang et al. (2024)				✓			
Mishra et al. (2025)				✓	✓		
Hsieh et al. (2023)					✓		
Huang et al. (2024)					✓	✓	
Cascante-Bonilla et al. (2024)					✓		
Zhu et al. (2025)					✓		
Kumar et al. (2024)						✓	
Gong et al. (2025)						✓	
Berasi et al. (2025)							✓
Total (25 papers)	9	3	8	3	7	5	1

Table 7: Where each analysis locates compositional failure. Obj. = the training objective or learning dynamics; Text = the text encoder; Vis. = the visual representation or encoder; Data = the training supervision; Eval. = the benchmark, scoring rule, or querying protocol; Cap. = a specific constituent capability such as counting or grounding; Geom. = the geometry of the embedding space. Marks reflect the primary claims of each paper; several papers argue for more than one location.

5.3 Where the Analyses Point

The analyses reviewed in Section 4.2 all ask the same question: what is responsible for compositional failure? They give different answers. Table 7 records where each one locates the problem. A paper may point to more than one place, and rows are ordered by the location each argues for most directly.

Two things stand out. The first is how spread out the answers are. The training objective and the visual representation attract the most attention, but the text encoder, the training data, the evaluation procedure, the constituent skills, and the geometry of the embedding space are each argued for by more than one paper, or argued for convincingly by one. The second is that these answers do not contradict each other. A model can be trained on an objective that never rewards structure, lose what structure it does encode in a pooled

text representation, and then be measured by a rule that fails to credit what survives. All three can be true at once.

That is what makes the picture difficult. Each account is supported by evidence, and each explains part of the failure, but nothing in the literature establishes how much of the total each one is responsible for. The studies are run on different benchmarks, with different models, under different scoring rules, so their results cannot be added up. The field has accumulated explanations without ranking them, and the reason is that a comparison capable of ranking them has not been carried out.

5.4 What the Mitigations Change

Table 8 compares the proposed methods along three lines: what part of the process each one changes, which parameters it updates, and whether the paper reports what the method does to abilities other than compositionality.

The first two columns show that the methods are more varied than the term "hard-negative training" suggests. Around half construct training examples, but the rest change the loss, the reward, the prediction order, the architecture, or nothing at all. The methods that update no parameters are worth noting as a group, because they cannot damage what the model already does, which makes them safe in a way the others are not.

The third column is the more challenging one. Compositional gains matter only if the model remains good at what it was already good at, and a method that improves a benchmark score while degrading retrieval has not clearly improved the model. Roughly a third of the methods reviewed here report this alongside their compositional results. The rest report the benchmark score alone, which means that for most of the literature it is not possible to say whether the improvement came at a cost. Where the cost is reported, it is often substantial, and this is discussed further below.

The fourth column records which benchmarks each method was tested on, and it explains why the methods are difficult to rank. There is no evaluation that all of them share. SugarCrepe and Winoground come closest, but even these are absent from a large number of the papers, and the methods aimed at specific axes are each tested principally on the benchmark introduced alongside them. Comparing two methods therefore usually means comparing numbers produced under different conditions.

5.5 Persistent Challenges

Three challenges emerge repeatedly across the reviewed literature. Each is reported by more than one study, and each complicates a claim the field otherwise takes for granted.

Gains do not transfer. Improving one compositional benchmark does not reliably improve another. For example, Miranda et al. (2024) find that results on a text-side benchmark and on their bidirectional one are uncorrelated: the best image-to-text model is not the best bidirectional model, and a plainly fine-tuned baseline can outperform methods trained specifically on hard negative captions once text-to-image retrieval is included. They also show this within a single controlled comparison, by training on hard negative captions and on hard negative images separately and observing that each leads on a different benchmark. Dumpala et al. (2024) report the same problem across formats rather than directions: succeeding at binary discrimination does not imply succeeding at a paraphrase test, because a model can learn to notice that a caption changed without learning whether the change mattered. If compositionality were one ability, this would not happen.

Improvements have costs, and the costs are usually unreported. Oh et al. (2024) show that hard-negative fine-tuning tends to weaken zero-shot classification and retrieval, and explain why: a hard negative is encoded almost identically to its positive in a single pooled vector, so the gradient that separates them disturbs the representation other tasks rely on. Peleg et al. (2025) report the same trade-off from the other side, finding that earlier methods improving compositional accuracy often degrade retrieval. Mishra et al. (2025) observe it in generative models, where an unfiltered hard-negative baseline matches their method on

Method	Acts on	Updates	Evaluated on	Gen. cap.
NegCLIP (Yuksekgonul et al., 2022)	Training data	Both encoders	ARO; retrieval, classification	✓
Cross-modal attn. (Tang et al., 2023)	Architecture	Added module	NCD/UNCD; VQA	×
Hard positives (Kamath et al., 2023a)	Training data	Both encoders	ATA, brittleness	×
VisMin (Awal et al., 2024)	Training data	Both; QLoRA	VisMin; COCO retrieval	✓
TROHN-IMG (Miranda et al., 2024)	Training data	Both encoders	BiVLC, SugarCrepe	×
CounterCurate (Zhang et al., 2024)	Training data	Both; MLLM	Flickr30k-Positions, SugarCrepe	×
SPARCL (Li & Li, 2025)	Data + loss	CLIP	VL-CheckList, SugarCrepe, SugarCrepe++	×
NegBench (Alhamoud et al., 2025)	Training data	CLIP	NegBench; retrieval	✓
SCRAMBLe (Mishra et al., 2025)	Training data	MLLM (pref. tuning)	Winoground, EqBen, COLA, ConMe; MM-Vet	✓
CLIC (Peleg et al., 2025)	Data + loss	Text encoder only	SugarCrepe++, Winoground; retrieval	✓
FSC-CLIP (Oh et al., 2024)	Loss	Both encoders	11 comp., 21 zero-shot, 3 retrieval tasks	✓
SPEC loss (Peng et al., 2024)	Loss	Both encoders	SPEC, ARO, EqBen; zero-shot	✓
Heatmap pretext (Zeng et al., 2024)	Loss	CLIP, ALBEF	ARPGrounding; grounding benchmarks	×
RL-Ground (Li et al., 2026)	Reward	Full model (RL)	ComPABench (ID and OOD)	×
IL-CLIP (Zheng et al., 2024)	Training dynamics	Both, from scratch	SugarCrepe; 18 classification datasets	✓
COGT (Parascandolo et al., 2025)	Prediction order	Decoder only	ARO, SugarCrepe, VL-CheckList, ColorSwap	×
SADE filtering (Ma et al., 2024)	Protocol	None	SADE	–
Example captions (Kumar et al., 2024)	Prompting	None	Compun	–
TTM (Zhu et al., 2025)	Inference	Pseudo-label tuning	Winoground, MMVP-VLM, SugarCrepe, WhatsUp	×
CECE (Cascante-Bonilla et al., 2024)	Inference	None	Winoground, EqBen	–
COCO-Tree (Sinha et al., 2025)	Inference	None	Winoground, EqBen, ColorSwap, SugarCrepe	–
GDE (Berasi et al., 2025)	Inference	None	CZSL benchmarks; group robustness	–

Table 8: Comparison of the proposed mitigations. "Acts on" is the part of the process the method changes; "Updates" is what is modified during training; "Evaluated on" lists the benchmarks the paper reports; "Gen. cap." records whether the paper reports the effect on zero-shot classification or retrieval alongside compositional results (✓ yes, × no), with "–" where the method updates no parameters and the question does not arise. Two further papers propose model-design (Hua et al., 2024) and data-design (Gong et al., 2025) recommendations without an implemented method, and so are not listed.

compositional benchmarks while measurably weakening general question answering. As Table 8 shows, most papers do not report this information at all, which means the trade-off has been documented where it was looked for and left unmeasured everywhere else.

The size of the gap depends on how it is measured. Two lines of work correct the measurement rather than the model, and they pull in opposite directions. Correcting the data lowers scores: Hsieh et al. (2023) show that models can exploit implausible or ungrammatical negatives without consulting the image, and Huang et al. (2024) show that negatives written without sight of the image can be rejected on plausibility alone, with scores falling substantially once this is fixed. Correcting the metric raises scores: Zhu et al. (2025) show that a conjunctive rule refuses to credit structure a model demonstrably has, and that a matching rule applied to the same predictions recovers a large amount of apparent capability. Both corrections are legitimate. Together they mean the reported gap is not a noisy estimate of one quantity but a mixture of at least three: real representational failure, artifacts in the data, and undercounting by the metric. Oh et al. (2024) add a fourth for any model that has itself been compositionally fine-tuned. No study measures the relative size of these components on common ground.

5.6 Synthesis

The comparisons above point to a consistent conclusion. That vision–language models are compositionally weak is not in doubt: every benchmark reviewed here places them below human performance, and on several the difference is large. What is in doubt is how weak, in what respect, and whether any particular method has helped.

Three reasons for this doubt have emerged. Measurement is not settled, so the same predictions can be scored as near-total failure or as reasonable competence depending on a rule that papers do not vary. Compositionality behaves as several abilities rather than one, so a result on any single benchmark does not generalize to the others. And methods are evaluated on non-overlapping sets of benchmarks, usually without reporting what they cost elsewhere, so they cannot be placed on a common scale.

None of these is a limitation of the models. All three concern how the field measures and reports, which means they can be addressed with the benchmarks, metrics, and diagnostic tools this literature has already produced. Section 6 takes up what that would involve.

6 Open Problems and Future Directions

The comparisons in Section 5 leave a number of questions that no reviewed paper answers. Five stand out, and each suggests work that could be carried out with the benchmarks and diagnostic tools the field already has.

Measurement is not settled. Section 5.2 showed that eight kinds of scoring rule are in use and that no benchmark reports its results under more than one. Since the rule alone can decide whether a model looks incompetent or capable (Zhu et al., 2025), comparisons across papers rest on an uncontrolled variable. The same applies to the size of the deficit itself, which is known to contain real representational failure, artifacts in the data, and undercounting by the metric, but never measured component by component on the same evaluation. Two things would help: reporting results under several scoring rules, which is cheap because most are computed from the same similarity scores, and a controlled study that varies the rule and the negative-generation procedure while holding the data fixed, so that the parts of the gap can be estimated rather than listed.

The visual side remains underexplored. Most benchmarks perturb only the caption, and the recent shift toward hard negative images and text-to-image retrieval (Miranda et al., 2024; Peng et al., 2024) is still bounded by the scale and fidelity of current image generators. Editing an image to change one relation while leaving everything else intact is difficult, and Li & Li (2025) find that uncontrolled synthetic variation introduces its own artifacts. Promising directions include generation methods that keep an edit local by construction rather than by filtering afterwards, and training objectives that strengthen the vision encoder

rather than the text encoder alone, since the persistent gap in the bidirectional setting shows that balanced, symmetric compositional understanding is still unsolved.

Composing versus the skills being composed. Most benchmarks report a single aggregate score, leaving it undetermined whether a failure reflects an inability to compose or a weakness in one of the skills being composed. The capability-annotated design of Gong et al. (2025) shows the value of separating the two, and its finding that counting is the dominant bottleneck, and does not improve with scale, suggests that discrete numerical reasoning under occlusion and clutter deserves attention as a target for supervision in its own right. Extending capability-level annotation to the two-dimensional benchmarks that dominate this survey, and to 3D-input models whose evaluation remains affected by input adaptation, would make this distinction measurable across the field rather than in a single benchmark.

Most evidence concerns one kind of model. Several of the reviewed papers note in their own limitations that their benchmark or method was evaluated only on contrastively trained models. The concentration is visible in Table 8, where most methods act on CLIP or a close relative, and it matters because the two architectural families do not fail in the same way. Tang et al. (2023) find that the attribute-binding mechanism they describe is specific to contrastive training and does not appear in autoregressive models, and Ma et al. (2024) show that generative models are scored by a rule that has a problem that contrastive scoring does not have. Conclusions established on one family therefore cannot be assumed to hold for the other. The papers that evaluate both (Awal et al., 2024; Zhang et al., 2024; Peng et al., 2024) show that this is practical rather than prohibitive, and extending existing benchmarks and mitigations to generative models would establish which of the field's findings describe compositionality itself and which describe a property of dual-encoder architectures.

Methods cannot be compared. Table 8 shows that the proposed mitigations share no common evaluation and that most do not report what their improvement costs in zero-shot classification or retrieval. Neither problem is difficult to fix. A small shared evaluation set, together with the convention that compositional results are reported alongside retrieval and zero-shot performance, would make the methods directly comparable and would show whether the trade-off documented by Oh et al. (2024) and Peleg et al. (2025) is general or particular to certain designs. The same argument applies to the two-sided criterion of Kamath et al. (2023a) and Dumpala et al. (2024): rejecting an altered meaning and accepting a preserved one are separate abilities, and reporting only the first leaves the second unmeasured.

How much is missing, and how much is inaccessible? The training-free results of Zhu et al. (2025), Cascante-Bonilla et al. (2024), and Berasi et al. (2025) raise the question of how much of the reported compositional gap is a representational deficit at all, as opposed to a failure to elicit structure the model already encodes. All three recover substantial gains without updating any parameter, one by correcting the matching procedure, one by restating the caption through entailments and contradictions, and one by decomposing embeddings along the geometry of the embedding space rather than linearly, and all three are bounded by what the vision encoder can resolve. Characterizing this ceiling, that is, establishing how far inference-time elicitation can go before visual representation becomes the binding constraint, would clarify which portion of the remaining gap fine-tuning is actually required to close. The question also has a practical side: a method that recovers existing structure cannot damage what the model already does, while one that rewrites the representation can.

Taken together, these are less a list of missing capabilities than a description of a field that can demonstrate a problem more reliably than it can measure it. That is a more tractable position than it may appear, since four of the five problems above concern how results are produced and reported rather than what the models are able to do.

7 Conclusion

Despite strong zero-shot and retrieval performance, vision-language models continue to exhibit severe and well-documented failures at compositional reasoning: binding attributes to the correct object, resolving

relations and spatial arrangements, understanding negation, and generalizing beyond memorized concept co-occurrences. This survey reviewed the key papers addressing this problem, organized into benchmarks that measure the gap with increasing control over potentially misleading factors, analyses that trace it to contrastive shortcut learning, a bottlenecked text encoder, and artifacts in evaluation data and metrics, and mitigation strategies ranging from hard-negative fine-tuning to training-free inference-time procedures. The comparison across this literature shows that the size of the compositional gap is not a fixed quantity: it shrinks substantially once benchmark artifacts and scoring-metric undercounting are corrected for, and reported gains from any single intervention frequently fail to transfer across retrieval directions, benchmarks, or to the model's original zero-shot and retrieval capability. At the current state of the field, compositionality in VLMs is best understood not as a single capability that can be summarized by a single score, but as a collection of related skills, including attribute binding, relational and spatial reasoning, negation, and bidirectional retrieval symmetry. Each presents distinct evaluation challenges and has prompted only partial solutions. Closing the remaining gaps will likely require advances not only in model training, but also in visual reasoning, in disentangling compositional ability from the underlying skills on which it depends, and in evaluation methodologies themselves.

Acknowledgments

The authors thank the teaching assistants of the System 2 course at Sharif University of Technology for their guidance throughout this project. Their suggestions of relevant papers and their feedback at each phase shaped both the selection of works reviewed here and the structure of the survey.

References

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29612–29622, 2025.

Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad Shahbaz Khan, and Abdul Wahab Muzaffar. Image colorization: A survey and dataset. *Information Fusion*, 114:102720, 2025.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. Vismin: Visual minimal-change understanding. *Advances in Neural Information Processing Systems*, 37:107795–107829, 2024.

Davide Berasi, Matteo Farina, Massimiliano Mancini, Elisa Ricci, and Nicola Strisciuglio. Not only text: Exploring compositionality of visual representations in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24917–24927, 2025.

Paola Cascante-Bonilla, Yu Hou, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. Natural language inference improves compositionality in vision-language models. *arXiv preprint arXiv:2410.22315*, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2236–2250, 2022.

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

Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Zhaokai Wang, Songze Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. SpaCE-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. *arXiv preprint arXiv:2506.07966*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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

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

Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024.

Irene Huang, Wei Lin, M. Jehanzeb Mirza, Jacob A. Hansen, Sivan Doveh, Victor Ion Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, Chuang Gan, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. Conme: Rethinking evaluation of compositional reasoning for modern vlms. In *Advances in Neural Information Processing Systems*, 2024.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. The hard positive truth about vision-language compositionality. *arXiv preprint arXiv:2305.01696*, 2023a.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4933–4944, 2023b.

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.

Sonal Kumar, Sreyan Ghosh, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. Do vision-language models understand compound nouns? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 519–527, 2024.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

Youngwan Lee, Soojin Jang, Yoorhim Cho, Seunghwan Lee, Yong-Ju Lee, and Sung Ju Hwang. Can vlms handle multi-hop compositional spatial reasoning? 2026.

Haoxin Li and Boyang Li. Enhancing vision-language compositional understanding with multimodal synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Tianle Li, Jihai Zhang, Yongming Rao, and Yu Cheng. Unveiling the compositional ability gap in vision-language reasoning model. *Advances in Neural Information Processing Systems*, 38:126574–126592, 2026.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Teli Ma, Rong Li, and Junwei Liang. An examination of the compositionality of large generative vision-language models. *arXiv preprint arXiv:2308.10509*, 2024.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. BiVLC: Extending vision-language compositionality evaluation with text-to-image retrieval. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. Scramble: Enhancing multimodal llm compositionality with synthetic preference data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6233–6243, 2025.

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained VLMs for improving vision-linguistic compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Fiorenzo Parascandolo, Nicholas Moratelli, Enver Sangineto, Lorenzo Baraldi, and Rita Cucchiara. Causal graphical models for vision-language compositional understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in CLIP with efficient fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

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

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15617–15629. IEEE, 2022.

Sanchit Sinha, Guangzhi Xiong, and Aidong Zhang. Coco-tree: Compositional hierarchical concept trees for enhanced reasoning in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2695–2711, 2025.

Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14333–14348, 2023.

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

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

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

Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn composable primitive concepts? *Transactions on Machine Learning Research*, 2023.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14141–14151, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

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

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13785–13795, 2024.

Yinglun Zhu, Jiancheng Zhang, and Fuzhi Tang. Test-time matching: Unlocking compositional reasoning in multimodal models. *arXiv preprint arXiv:2510.07632*, 2025.