Can VLMs Actually See and Read? A Survey on Modality Collapse in Vision-Language Models

Mong Yuan Sim[1,2] **Wei Emma Zhang**[1] **Xiang Dai**[2] **Biaoyan Fang**[3]
[1]The University of Adelaide [2]CSIRO Data61 [3]Oracle
{mongyuan.sim;wei.e.zhang}@adelaide.edu.au

Abstract

Vision-language models (VLMs) integrate textual and visual information, enabling the model to process visual inputs and leverage visual information to generate predictions. Such models are demanding for tasks such as visual question answering, image captioning, and visual grounding. However, some recent work found that VLMs often rely heavily on textual information, ignoring visual information, but are still able to achieve competitive performance in vision-language (VL) tasks. This survey reviews modality collapse analysis work to provide insights into the reason for this unintended behavior. It also reviews probing studies for fine-grained vision-language understanding, presenting current findings on information encoded in VL representations and highlighting potential directions for future research.

1 Introduction

Integration of information from multiple sensory modalities, such as language and vision is crucial in forming a cohesive understanding of the world. Humans naturally combine sensory inputs in a way that balances and enhances the contributions of each modality. This is called *cross-modal integration*, which allows humans to interpret complex environments effectively and make inferences that go beyond any single information source (McGurk and MacDonald, 1976; Shams and Seitz, 2008).

Despite the clear advantages of cross-modal integration observed in human cognition, many vision-language models (VLMs) struggle with modality collapse problems and fail to achieve a similar balance (Jabri et al., 2016; Goyal et al., 2018; Frank et al., 2021). This problem arises when a model fails to utilize one modality (modality collapse) and only relies on another (modality dominance). Modality collapse is when an unimodal model achieves similar accuracy on a vision-language task compared to a multimodal model, showing the other modality is not fully utilized (Javaloy et al., 2022; Parcalabescu and Frank, 2023; Liang et al., 2024; Gapp et al., 2025). Modality collapse can impact the reliability of VLMs, especially in tasks requiring a fine-grained understanding of both vision and text. Figure 1 shows examples where even powerful VLMs fail on simple tasks due to modality collapse.

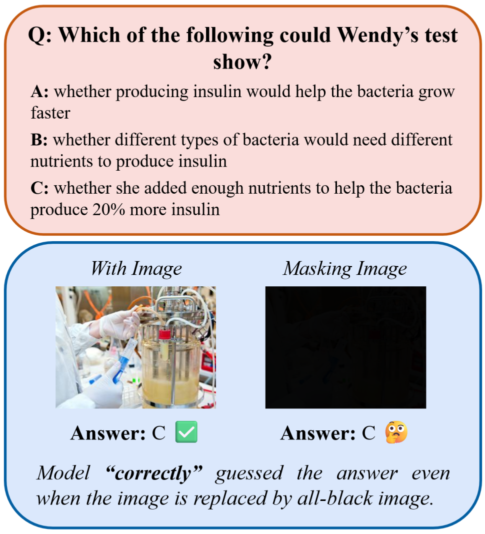

Figure 1: An example of modality collapse, where a unimodal VQA model still selects the "correct" option, as if it could see and read the image, even when the input is an all-black image.

Previous works mainly focus on improving the model's performance and robustness through debiasing (Berg et al., 2022; Si et al., 2023; Seth et al., 2023), increasing model size (Dehghani et al., 2023), and using more training data (Zhai et al., 2022). Nevertheless, the extent to which VLMs utilize vision and language modalities and their limitations remain unclear. To guide further research in VLMs, we collect and piece together existing knowledge about modality collapse in VLMs to complete the puzzle, answering the following research questions.

Findings of the Association for Computational Linguistics: ACL 2025, pages 24452–24470
July 27 - August 1, 2025

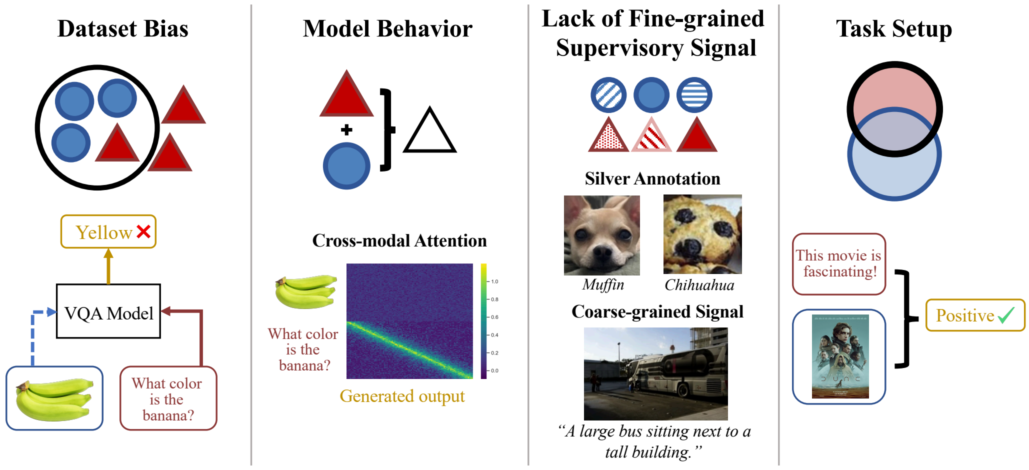

Figure 2: The contributing factors to **modality collapse** in the VLMs, including (1) *dataset bias*: that can cause VLMs to ignore visual input, (2) *model behavior*: where the model unintentionally learned to utilize one modality more than the other, (3) *lack of fine-grained supervisory signal*: existing pretraining paradigms often rely on automatically annotated silver annotations and coarse-grained signal and (4) *task setup*: some tasks are meant to use one modality only by nature.

RQ1: Do VLMs Effectively Utilize Both Modalities When Relevant? State-of-the-art VLMs demonstrate strong performance on various VL tasks (OpenAI et al., 2024; Liu et al., 2023b). However, questions persist about the extent to which these models genuinely utilize both modalities when relevant. Previous studies showed that, in practice, text modality often dominates, leading to concerns about whether these models exhibit true vision-language understanding (Cao et al., 2020; Zhu et al., 2022). In addition, VLMs can exploit textual bias in the dataset, neglecting the image input (Jabri et al., 2016; Goyal et al., 2019; Srinivasan and Bisk, 2022). Comparing performance between unimodal models and multimodal models does not reflect the utilization of different modalities. Therefore, it is crucial to have methods to quantify the modality contribution and assess the existence of cross-modal interaction.

RQ2: What Are (Not) Encoded in VL Representations? Following RQ1, we aim to further understand what information is encoded in VL representations and what is not. As vision modality collapse is very common in VLMs (Goyal et al., 2019; Frank et al., 2021; Zhu et al., 2022), we hypothesize that these models struggle to encode fine-grained information and hence perform poorly on tasks requiring vision-language compositionality.

Based on the two research questions above, this survey systematically reviews the contributing factors to modality collapse and information encoded by VLMs. Our key contributions are as follows:

- This paper comprehensively reviews recent advancements investigating modality collapse and dominance in VLMs, providing insights into contributing factors of modality collapse.
- It categorizes the information encoded in VL representations into three distinct dimensions: linguistic semantics, visual content, and vision-language compositional, providing the first taxonomy for understanding VL representations' capabilities and limitations.
- Building on these findings, this paper proposes actionable future directions to allow VLMs to utilize both text and vision modalities, generating more reliable predictions.

Related Surveys Several surveys have been conducted to review VLMs (Du et al., 2022; Long et al., 2022), multimodal models[1] (Uppal et al., 2022; Xu et al., 2023; Liang et al., 2024), large multimodal models (Yin et al., 2024; Caffagni et al., 2024; Wu et al., 2023) and hallucination issue in large multimodal models (Bai et al., 2024). To the best of our knowledge, our survey is the first one that reviews the utilization of vision and language modalities in VLMs and their limitations in encoding fine-grained information.

[1]The scope of this paper is vision-and-language only. When the term "multimodal" is used, it refers to vision and language or combinations that also include other modalities.

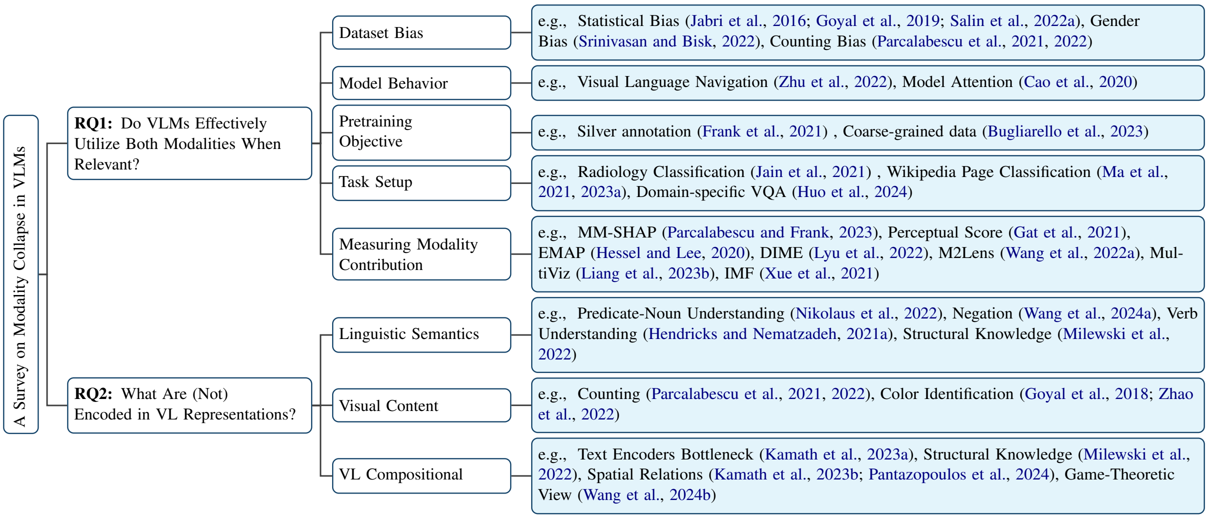

Figure 3: Taxonomy of modality collapse contributing factors and information encoded in vision-language (VL) representations.

2 Preliminary

We group existing vision-language models (VLMs) into four architectural categories based on their vision-language fusion strategies: single-stream, dual-stream, dual-encoder, and large vision-language models (LVLMs).

Single-stream VLMs refer to VLMs that process text and image input with a single Transformer encoder. Examples in this category include VL-BERT (Su et al., 2020), VisualBERT (Li et al., 2019a), UNITER (Chen et al., 2020), Unicoder (Li et al., 2020a) and Oscar (Li et al., 2020b). This design is also referred to as early fusion, as text and visual inputs are combined at the input level.

Dual-stream VLMs process text and image inputs independently using dedicated encoders, and then fuse their representations via a multimodal fusion module (e.g., co-attention). The fusion is performed in a shared transformer designed to jointly reason over both modalities. Early dual-stream VLMs include BERT-based VLMs such as ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019).

LVLMs or *Large Vision-Language Models* extend pretrained large language models by incorporating visual inputs through an adapter (e.g., MLP or Q-former). Some LVLMs (e.g., LLaVA (Liu et al., 2023b)) directly project image embeddings into the language model's input space, and some use a more complex network (e.g., Q-former in BLIP-2 (Li et al., 2023)) to perform lightweight vision-language fusion before passing the result to the language model. Unlike dual-stream models, LVLMs avoid deep co-attentional fusion and rely on the language model to generate outputs from injected visual context.

Dual Encoder Models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) encode images and text separately, aligning them in a shared embedding space via contrastive learning. Unlike other architectures, they perform no cross-modal fusion, making them efficient for retrieval tasks but less suited for generation or fine-grained reasoning.

3 RQ1: Do VLMs Effectively Utilize Both Modalities When Relevant?

Takeaway Message: *No, modality collapse often happens in VLMs and text modality often dominates*. Though VLMs showed outstanding performance on several VL tasks when tested on benchmarks, little is known about whether they really "see" and "read" the input image. In this section, we divide this research question into two parts: contributing factors to modality collapse, and methods to measure the contribution of different modalities.

3.1 Contributing Factors to Modality Collapse

In this subsection, we systematically review works that investigate the contribution of vision and language modalities, grouped by their findings which lead to modality collapse in VLMs.

3.1.1 Dataset Bias

VLMs are often trained on datasets where textual information dominates over visual content, due to statistical bias that hinders the effective utilization of visual modality. This over-representation of text can lead models to rely heavily on the textual modality while ignoring visual cues, even for tasks requiring detailed visual information. Studies in this category examine statistical bias in the datasets that can cause modality collapse in VLMs. Statistical bias in datasets refers to the disproportionate representation of certain features or categories, such as gender biases, arising from high-frequency occurrences. For example, VQA task requires VLMs to obtain answers from images. However, earlier work finds VQA models can exploit statistical textual bias and cause modality collapse, where the model can achieve competitive performance without accessing the image input (Jabri et al., 2016) and rely on the first few text tokens in the question (Agrawal et al., 2016).

Parcalabescu et al. (2021) probe VLMs ability to count and reveal that VLMs struggle with counting, often defaulting to predicting common quantities in the datasets, rather than accurately interpreting visual information. For example, VLMs favor frequent numbers such as "two" (predicted by the model 51% of the time), while larger numbers are predicted less frequently, showing poor generalization of VLMs in counting tasks. Similar to counting tasks, Salin et al. (2022b) find that VLMs struggle in understanding size and position information, where the models rely heavily on text input and learned textual bias. They further show that fine-tuning on specially crafted data does not lead to better performance. Srinivasan and Bisk (2022) shows that VL-BERT exhibits notable gender biases, where stereotypical assumptions override the actual visual inputs. Statistical bias not only causes vision modality to collapse, but also makes VLMs less reliable and poses safety issues in real-world applications.

3.1.2 Model Behavior

Previous studies find that some design choices, though unintentional, could lead the VLMs to rely more heavily on text, contributing to modality collapse This section explores how model behavior leads to modality collapse.

Most existing pretrained VLMs have a special `[CLS]` token which absorbs information from text and vision modalities through self-attention. Cao et al. (2020) find that `[CLS]` token has a higher attention on text over image input, absorbing more information from text modality for VQA task in the general domain.

Zhu et al. (2022) probe VLMs for Visual Language Navigation (VLN) task to assess the importance of language and vision modalities. The authors find that masking text tokens caused a sharp performance drop while masking all visual tokens did not. This contradicts the definition of VLN task, where vision input should be the primary source of information to generate output.

Recent works have also explored different model behaviors contributing to the ineffectiveness of cross-modality interaction in LVLMs. Zhang et al. (2024a); Kaduri et al. (2024) find that LVLMs decoder attends to irrelevant tokens, leading to ineffective visual input processing. Zhu et al. (2024) highlights cross-modality knowledge conflicts, where inconsistencies between vision encoder and language model lead to misalignment and suboptimal fusion of multimodal information. Additionally, Zhang et al. (2024b) identifies a conceptual mismatch problem caused by contrastive learning training paradigm, where text-image pairs may not always align semantically.

3.1.3 Lack of Fine-grained Supervisory Signal

The supervisory signal in pretraining plays a pivotal role in shaping VLMs performance. In VL setting, fine-grained pretraining is to train a model to capture more detailed local information within the image and map it to the corresponding text segment, while coarse-grained pretraining only aim to align the whole image with its corresponding text description.

The importance of fine-grained pretraining signal is shown in Bugliarello et al. (2023), where they find that modeling objects has more impact than increasing data scale. The authors find that VLM trained on a smaller size fine-grained dataset, e.g., X-VLM$_{4M}$ (Zeng et al., 2022), outperform BLIP$_{129M}$ (Li et al., 2022), which is trained on 129M coarse-grained data points. Besides, X-VLM is trained on image region and text matching, and bounding box prediction task, forcing it to learn visual grounding by aligning text descriptions with specific object regions in an image. In contrast, BLIP is pretrained on image-text matching only, without enforcing the connections between image regions and text segments. These findings highlight the importance of fine-grained training objective, which can shape VLMs' ability in utilizing visual input, instead of relying on text cues only.

Many BERT-based VLMs (e.g., VL-BERT (Su et al., 2020) and ViLBERT (Lu et al., 2019)) use silver annotations from Faster-RCNN (Ren et al., 2016) as training data. However, Frank et al. (2021) observe that these silver annotations are not reliable when compared to the gold labels with only 38% agreement. This raises concerns about VLMs trained with such noisy supervision signal can truly develop fine-grained understanding from visual information, or if they learn to rely on linguistic cues, which lead to modality collapse.

3.1.4 Task Setup

Multimodal setting is first proposed to address the limitations of unimodal models, by enabling models to process and integrate information from multiple modalities. However, not all task setups can benefit from multimodal settings. Certain tasks inherently require multiple modalities as inputs (e.g.,VQA requires image input and corresponding text questions), while others are initially defined as unimodal tasks, where additional modalities serve as supplementary information (e.g., multimodal summarization).

Ma et al. (2023a) conducted an annotation study on Japanese Wikipedia text classification task. The dataset is curated from Wikipedia pages and the task aims to classify them into corresponding named entity classes Ma et al. (2021). Human annotators find that images tend to be distracting and misleading. An annotation study in the radiology domain (Jain et al., 2021) also showed a similar finding. Different groups of radiologists are asked to label radiology images and radiology reports and compare them against ground truth. There is a significant disagreement between labels from these two groups of annotators. One of the reasons mentioned by Jain et al. is the difference in modality-specific context: radiologists labeling reports have access to clinical history and additional contextual information, while those labeling images rely solely on visual cues. This highlights how different input modalities, such as text and image modalities, can lead to varying interpretations, even among experts. Consequently, the choice of modality, either vision or language, can influence the labeling outcomes, and the effectiveness of each modality often depends on the specific task and setup.

Huo et al. (2024) conduct a neuron-level analysis to understand the utilization of vision modality in LVLMs. They concluded that deactivating domain-specific multimodal neurons in some domains (e.g., medical and auto-driving domain) does not cause a sharp decrease in overall performance, showing that vision inputs are not required for those tasks.

Discussion It is essential to highlight that the factors contributing to modality collapse are not independent. For instance, noisy pretraining dataset can cause VLMs to exploit textual bias and task setup that do not need visual information can cause a model to rely more on text input (or vice versa). Therefore, it is important to ensure that the pretraining and finetuning dataset are unbiased or de-biased and the task itself needs both visual and textual input, in order to unlock the capability of VLMs in utilizing both modalities.

3.2 Measuring Modality Contributions

One simple way to demonstrate the usefulness of different modalities is to compare the performance of unimodal models with multimodal models (Wang et al., 2022c; Hu et al., 2023; Li et al., 2024). The intuition is, if a multimodal model that uses both text and image inputs outperforms a unimodal model (typically using text only) on the test set, it is generally assumed that images are effectively contributing to the task. However, this simple comparison overlooks many other factors, such as model size and dataset bias (Yogatama et al., 2015; Dodge et al., 2019; Hessel and Lee, 2020). For instance, a unimodal modal can be "upgraded" to a multimodal model by adding a projection layer to project vision representation obtained from pretrained vision encoder to the language model representation space. However, even when the vision representation is irrelevant, the performance might still be improved due to the increase of trainable parameters, given the language models are of the same size (Du et al., 2022; Long et al., 2022).

Therefore, methods for quantifying modality contribution and cross-modal interaction are explored to assess the effectiveness of different modalities in a more controlled setup. In this section, we review methods for measuring how different modalities contribute to downstream prediction and cross-modal interaction within VLMs.

Modality contribution refers to *the extent to which a given modality influences model predictions* (Parcalabescu and Frank, 2023; Liang et al., 2023b). Parcalabescu and Frank (2023) propose a contribution measure, MM-SHAP, inspired by cooperative game theory (Shapley, 1953; Lundberg and Lee, 2017). MM-SHAP randomly mask pairs of text and image tokens, computes the output, and measures the change in probabilities compared to those obtained with the original inputs. MM-SHAP and its extension, CC-SHAP (Parcalabescu and Frank, 2024), can be applied to encoder-only VLMs and VLMs with a decoder, respectively. Gat et al. (2021) introduce *Perceptual Score* to assess the degree to which a model relies on different subsets of the input features (i.e., a combination of modalities). After training the classifier, they permute the features of a modality across test samples and observe the impact on performance. A significant drop in accuracy indicates a high contribution from the permuted modality.

Cross-modal interaction refers to *how different modalities relate with each other and potentially create new information that unimodal cannot achieve* (Liang et al., 2023b). This line of work aims to disentangle cross-modal interactions from multimodal models and observe changes in output logits or overall performance. Hessel and Lee (2020) introduce EMAP, a formal definition and method to measure cross-modal interactions with statistical non-additive interactions. That is, a function truly learns cross-modal interaction when it cannot be decomposed into two separate subfunctions that each process a single modality independently and then simply combine their results. This means the function must process the different modalities in an interconnected way, rather than handling each modality in isolation and merely adding their individual contributions. DIME (Lyu et al., 2022) extends EMAP and LIME[2] (Ribeiro et al., 2016) to enable feature visualization and explanation for each data instance. MultiViz (Liang et al., 2023b) incorporates EMAP's ability to disentangle unimodal and cross-modal contributions globally, and DIME's feature visualization for disentangled representations locally, while introducing a novel second-order gradient approach that can scale to more than two modalities. It uses a sparse linear model to understand how features are composed for final predictions. Similarly, Wang et al. (2022a) proposes M2Lens, an interactive multimodal sentiment analysis system. M2Lens uses SHAP values to group inputs into three groups: *dominance*, *complement*, and *conflict*. This categorization enables the visualization of connections between modalities and tokens.

Some interpretability methods can be used to understand the utilization of vision and language input. For instance, neuron-level interpretability methods can show VLMs' sensitivity to vision and language representations (Huo et al., 2024; Dai et al., 2022; Pan et al., 2024). Techniques such as logit lens (Daujotas; Neo et al., 2024), gradient-based (Rajabi and Kosecka, 2024), attention-based (Jiang et al., 2024; Chefer et al., 2021) visualization, and causal tracing tools (Palit et al., 2023; Basu et al., 2024) enable tracing information flow within Transformer models, revealing how visual and textual representations influence final predictions. Although these interpretability methods do not directly quantify and measure modality contribution, they can enhance our understanding of cross-modal interactions and help diagnose modality collapse or dominance.

Discussion Existing modality contributions and cross-modal interactions metrics have their strengths and limitations. Perturbation-based methods like MM-SHAP and EMAP are computationally expensive as they need to compute all possible pairs of inputs. They are more suitable for showing the overall modality contribution for a dataset. Metrics like MultiViz and DIME are more ideal for visualization purposes, showing the important parts of the inputs.

In addition, we note that almost all modality contribution and cross-modal interaction measures are perturbation-based. They compare the outputs from paired and unpaired text-image inputs to quantify the degree of contribution and interaction. While effective in capturing some aspects of cross-modal interaction and modality contribution, this paradigm poses challenges. For example, real-world datasets can contain various types of statistical bias (e.g., the word "dog" frequently co-occurs with images featuring grass). A perturbation-based method might remove text input ("dog") to assess the model's reliance on visual information. As the model has learned a spurious correlation between "dog" and "grass" during training, it might still perform well using the grass in the image background as a hint. Hence, these measures might underestimate the contribution of the permuted modality.

[2]LIME is a perturbation-based method that works for unimodal input only. It first breaks unimodal input into different parts, and randomly modifying these parts multiple times to see how each change affects the model's output. It then trains a linear model to shows which part of the input is the most important for the model's decision.

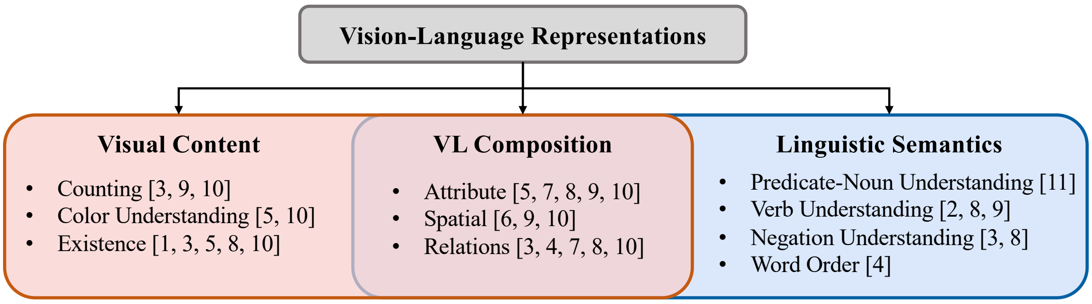

Figure 4: Fine-grained vision-language probing tasks, grouped into **visual content**, **vision-language composition**, and **linguistics semantic**. Benchmark datasets that have subset to probe these information are as labeled: [1] FOIL IT! (Shekhar et al., 2017), [2] SVO (Hendricks and Nematzadeh, 2021b), [3] VALSE (Parcalabescu et al., 2022), [4] Winoground (Thrush et al., 2022), [5] VL-Checklist (Zhao et al., 2023), [6] Visual Spatial Reasoning (VSR) (Liu et al., 2023a), [7] ARO (Yuksekgonul et al., 2023), [8] CREPE (Ma et al., 2023b), [9] EQBEN (Wang et al., 2023) and [10] MMVP (Tong et al., 2024), [11] Predicate-Noun (Nikolaus et al., 2022). For full description of these benchmark datasets, please refer to Appendix B. Examples for each category are shown in Figure 5.

4 RQ2: What Are (Not) Encoded In VL Representation? Through The Lens of Probing Studies

Takeaway Message: *VLMs can encode basic linguistic structure, simple fine-grained information (one-object setting), but fail to encode rich compositional information*[3]. Following RQ1, we see that VLMs are often dominated by text modality, due to textual bias, task setup, and lack of fine-grained training objective. Studies reviewed in RQ1 reveal the limitations of VLMs, showing that VLMs perform relatively well on coarse-grained tasks but fail to utilize visual information on fine-grained tasks.

In this section, we aim to understand what is encoded in VL representation, categorized into three categories: i) linguistic semantics, ii) visual content, and iii) VL compositional.

4.1 Linguistic Semantics

As VLMs are trained by aligning images and their corresponding text descriptions , *do they learn and encode linguistic semantics?* Hendricks and Nematzadeh (2021a) collected a benchmark dataset for verb understanding and tested verb understanding in VLMs like UNITER, ViLBERT, and LXMERT. Their results show that verbs are harder than subjects and objects and that models perform badly identifying negative captions.

Analysis by Ma et al. (2022) reveals that VLMs have a preference on visual tokens. The model learns to match the visual token in the caption to the corresponding image and discard global semantics. Milewski et al. (2022) show that multimodal BERT models encode less structural grammatical knowledge in the text embeddings, compared to text-only BERT.

Nikolaus et al. (2022) manually curate a dataset to test VLMs predicate-noun understanding. Results show that LXMERT and UNITER are among the best-performing models, while CLIP performs worse. The authors hypothesize that this is due to the pretraining objective, as LXMERT and UNITER have multimodal pretraining objectives, in addition to image-text matching.

Compared to pretrained VLMs, Wang et al. (2024a) show that LVLMs showed a better understanding of negation and triplet relationships (subject, verb, object), though still underperform on spatial relationship and compositional aspects (noun and attributes) which will be discussed later in Section 4.3.

4.2 Visual Content

When we use a VLM, we expect the model to really "see" an input image and provide a response based on the query. However, there are research works that show VLMs suffer from simple tasks that require visual perception only like counting (Seguí et al., 2015; Kamath et al., 2023a) and color identification. Parcalabescu et al. (2021) show that pretrained VLMs could not count and exploit statistical bias in the training dataset. Their follow-up work shows the same finding on their newly proposed benchmark dataset VALSE (Parcalabescu et al., 2022).

[3]Bexte et al. (2024) combine publicly available probing datasets into a unified benchmark dataset. Our review in this section aims to provide insights into when and why VLMs fail on such probing tasks, instead of providing an exhaustive list of all possible probing tasks.

Another simple yet underperformed task in VLMs is color identification. Similar to counting tasks, VLMs often exploit statistical bias in the training dataset rather than faithfully encoding fine-grained information (Zhao et al., 2022; Akula et al., 2024). For example, when asked "What color is the banana?", a model might answer "yellow" without looking at the image input (Goyal et al., 2019).

Contrary to previous findings, Salin et al. (2022a) find that image input is utilized by VLMs, by showing that mismatched image-text pairs lead to significant degrade in model performance. However, it is worth noting that they do not control dataset bias in the probing dataset. In particular, the probing dataset used includes only unambiguous colors, such as blue, red, and black, which may introduce bias to their findings.

4.3 Vision-Language Compositional

Vision-language compositionality refers to the ability of VLMs to understand components that form visual and textual information. It allows the model to distinguish between "the man is eating the steak" and "the steak is eating the man". This requires a VLM to encode both linguistic semantics and visual content into the VL representation, in order to recognize the presence of both a man and a steak, and correctly determine the relationship between them (i.e., who is performing the action and who is receiving it). This is essential for challenging downstream tasks like VQA, visual-language navigation (VLN), and image captioning. However, VLMs often take shortcuts by exploiting text input and do not utilize visual input.

A number of benchmark datasets have been created to understand VL compositionality in VLMs, such as: FOIL IT! (Shekhar et al., 2017), Winoground (Thrush et al., 2022), VALSE (Parcalabescu et al., 2022), and EVil-Probe (Bexte et al., 2024). For a full description of these benchmark datasets, please refer to Appendix B.

Kamath et al. (2023a) find CLIP (Radford et al., 2021) failed to encode compositional information, such as spatial information and relations. Parallel research supports this finding and further reveals that text encoders preserve compositional information better than vision encoders, a result that contradicts intuition (Milewski et al., 2022; Alper et al., 2023; Wang et al., 2024b).

Most recently, Hsieh et al. (2023) proposed the SugarCrepe benchmark dataset to evaluate VL compositionality and showed that many benchmark datasets (e.g., CREPE (Ma et al., 2023b), ARO (Yuksekgonul et al., 2023), VL-Checklist (Zhao et al., 2023)) are hackable, as they used a rule-based method to generate negative pairs, which can introduce unintentional biases, where the model can easily distinguish negative text caption, without truly understanding the image input. To reduce such biases, the authors generate hard negatives by using LLMs with human validation. Experimental results on SugarCrepe suggest that existing VLMs perform well on object recognition, but not on composing attributes and relations.

Discussion Many works attempt to create harder benchmark datasets, covering more visual patterns that previous works have missed (e.g., counting, position, attributes etc). They all point to the same conclusion: VLMs perform well on conventional VL tasks but fail on tasks that require VL compositionality (Parcalabescu et al., 2022; Hsieh et al., 2023; Zeng et al., 2024). To some extent, it is useful to finetune a pretrained VLM on datasets to improve VL compositionality. However, it is more effective to address this issue during the pre-training stage to ensure that VL compositionality generalizes better to unseen data. We discuss more actionable directions in Section 5.

5 Future Direction

A review of existing studies shows that modality collapse is common in VLMs. Although coarse-grained visual information (e.g., distinguishing paired image-text samples from unmatched pairs) is encoded into VL representations and benefits downstream tasks, current VLMs still struggle to encode fine-grained information (e.g., spatial relationships and attributes) and handle VL compositionally. To address these limitations and unlock the full potential of VLMs, we believe there are several future research directions.

Analysis on Modality Contribution and Cross-Modal Interaction Most methods reviewed in Section 3.2 rely on the perturbation of text-image pairs to measure the modality contribution and cross-modal interaction. Although these methods have intriguing model-agnostic features, they are likely affected by dataset bias and do not reflect real modality contribution and cross-modal interaction within a multimodal model. We believe one promising research direction is to develop a model-specific method that can directly analyze internal model representations and feature space (Huo et al., 2024), in order to quantify modality interaction. Another research direction is to develop a dataset-centric and model-agnostic metric, to evaluate the multimodal complexity of a task based on a dataset. Such metrics would help to determine the extent to which a task requires multimodal inputs. These two research directions disentangle the factors of the dataset and model behavior. Addressing these gaps is critical in mitigating modality imbalance and enhancing the robustness of VLMs.

Curation of Fine-Grained Probing Datasets Probing encoded information in VL representations is essential in understanding the capabilities and limitations of VLMs. However, most existing benchmarks rely on simple rules (e.g., syntactic modifications or basic attribute swaps), which may not truly assess the capabilities of VLMs.For instance, Hsieh et al. (2023) finds many existing datasets (e.g., CREPE (Ma et al., 2023b), ARO (Yuksekgonul et al., 2023)) contain samples that do not make logical sense and captions that contain obvious grammatical errors, which make it easy for VLMs to make the correct selection, even without accessing image inputs. In addition to conducting more controlled experiments to rule out these factors, future research could focus on building more rigorous datasets that minimize artifacts and biases to ensure model performance reflects genuine multimodal understanding, rather than reliance on spurious correlations.

Enhacing VLM Training Training VLMs is a data-hungry process, often relying on automatically scraped image-text pairs and auto-generated annotations, which may contain significant noise and often lack fine-grained information. This leads to significant statistical bias and unintended behavior, such as exploring non-visual attributes for concepts (Alper et al., 2023), and exploitation of statistical bias (Jabri et al., 2016). Future work should focus on developing high-quality training datasets with better text-image alignment (Peng et al., 2024) and fine-grained annotations. Simultaneously, pretraining objectives that explicitly encourage fine-grained understanding, such as using predicting bounding box and image region-text matching as objective (Zeng et al., 2022). As highlighted in (Bugliarello et al., 2023), while dataset scale contributes to generalization, it is the choice of pretraining objective that determines how a model balances its use of visual and textual modalities. Among the papers we reviewed, no model architecture stands out as the best solution to the problem of modality collapse.

6 Conclusion

This survey explores the issue of modality collapse in VLMs, highlighting the tendency of existing models to rely more on text input than visual information. We review factors that may contribute to this imbalance, including dataset bias, model behavior, pretraining objectives, and task setup. The issue of modality collapse also underscores the importance of properly evaluating cross-modal interactions, particularly in assessing whether VLMs genuinely and effectively utilize multimodal inputs. This paper is the first systematic review of modality collapse in VLMs. We hope this survey will facilitate further research in this area.

Limitations

This survey paper aims to understand whether visual input is being utilized in VLMs and what information is encoded into VL representation. We do not aim to cover all possible model architectures, tasks, and analyses in VLMs. Instead, we only focus on the issue of modality collapse, particularly the tendency of VLMs to rely more heavily on textual input than visual input. In addition, though related to VLMs' robustness, this survey does not extensively address robustness in VLMs. However, it is worth noting that the problem of modality collapse can be treated as a subtopic of robustness, emphasizing how multimodal inputs impact model performance and behavior.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Arjun Reddy Akula, Garima Pruthi, Inderjit S Dhillon, Pradyumna Narayana, Sugato Basu, and Varun Jampani. 2024. PRISM: A new lens for improved color understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1659–1670, Miami, Florida, US. Association for Computational Linguistics.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. 2023. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024. Understanding information storage and transfer in multi-modal large language models. *Preprint*, arXiv:2406.04236.

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, Online only. Association for Computational Linguistics.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2024. EVil-probe - a composite benchmark for extensive visio-linguistic probing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6682–6700, Torino, Italia. ELRA and ICCL.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, Toronto, Canada. Association for Computational Linguistics.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 565–580, Berlin, Heidelberg. Springer-Verlag.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396, Los Alamitos, CA, USA. IEEE Computer Society.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. *UNITER: UNiversal Image-TExt Representation Learning*, page 104–120. Springer International Publishing.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Gytis Daujotas. Case study: Interpreting, manipulating, and controlling clip with sparse autoencoders.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. 2023. Scaling vision transformers to 22 billion parameters. *Preprint*, arXiv:2302.05442.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5436–5443. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christian Gapp, Elias Tappeiner, Martin Welk, Karl Fritscher, Elke Ruth Gizewski, and Rainer Schubert. 2025. What are you looking at? modality contribution in multimodal medical deep learning methods. *Preprint*, arXiv:2503.01904.

Itai Gat, Idan Schwartz, and Alexander Schwing. 2021. Perceptual score: what data modalities does your model perceive? In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2018. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127(4):398–414.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vision*, 127(4):398–414.

Lisa Anne Hendricks and Aida Nematzadeh. 2021a. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Lisa Anne Hendricks and Aida Nematzadeh. 2021b. Probing image-language transformers for verb understanding. *CoRR*, abs/2106.09141.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *ArXiv*, abs/2306.14610.

Jinpeng Hu, Zhihong Chen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2023. Improving radiology summarization with radiograph and anatomy prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12066–12080, Toronto, Canada. Association for Computational Linguistics.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6816, Miami, Florida, USA. Association for Computational Linguistics.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *Computer Vision – ECCV 2016*, pages 727–739, Cham. Springer International Publishing.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A. Young, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. 2021. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, ACM CHIL '21, page 105–115. ACM.

Adrian Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9938–9964. PMLR.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2019. Action genome: Actions as compositions of spatio-temporal scene graphs. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10233–10244.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2024. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*.

Omri Kaduri, Shai Bagon, and Tali Dekel. 2024. What's in the image? a deep-dive into the vision of vision language models. *Preprint*, arXiv:2411.17491.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023b. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*, page 121–137. Springer International Publishing.

Yong-Lu Li, Liang Xu, Xijie Huang, Xinpeng Liu, Ze Ma, Mingyang Chen, Shiyi Wang, Haoshu Fang, and Cewu Lu. 2019b. Hake: Human activity knowledge engine. *ArXiv*, abs/1904.06539.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Russ R Salakhutdinov, and Louis-Philippe Morency. 2023a. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pages 27351–27393. Curran Associates, Inc.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023b. Multiviz: Towards visualizing and understanding multimodal models. In *International Conference on Learning Representations*.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, 56(10).

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

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

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

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. 2022. Vision-and-language pretrained models: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5530–5537. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. *ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*. Curran Associates Inc., Red Hook, NY, USA.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 455–467, New York, NY, USA. Association for Computing Machinery.

Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, and Timothy Baldwin. 2021. On the (in)effectiveness of images for text classification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 42–48, Online. Association for Computational Linguistics.

Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, and Timothy Baldwin. 2023a. On the effectiveness of images in multi-modal text classification: An annotation study. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(3).

Zheng Ma, Shi Zong, Mianzhi Pan, Jianbing Zhang, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2022. Probing cross-modal semantics alignment capability from the textual perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5739–5749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Victor Milewski, Miryam de Lhoneux, and Marie-Francine Moens. 2022. Finding structural knowledge in multimodal-BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5658–5671, Dublin, Ireland. Association for Computational Linguistics.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2024. Towards interpreting visual information processing in vision-language models. *Preprint*, arXiv:2410.07149.

Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. 2022. Do vision-and-language transformers learn grounded predicate-noun dependencies? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1538–1555, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. 2023. Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP . In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2848–2853, Los Alamitos, CA, USA. IEEE Computer Society.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024. Finding and editing multi-modal neurons in pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1012–1037, Bangkok, Thailand. Association for Computational Linguistics.

Georgios Pantazopoulos, Alessandro Suglia, Oliver Lemon, and Arash Eshghi. 2024. Lost in space: Probing fine-grained spatial understanding in vision and language resamplers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 540–549, Mexico City, Mexico. Association for Computational Linguistics.

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

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2024. Do vision & language decoders use images and text equally? how self-consistent are their explanations?

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. 2024. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *CVPR*.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhi Ding, Scott D. Cohen, Quan Tran, and Abhinav Shrivastava. 2021. Learning to predict visual attributes in the wild. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13013–13023.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *European Conference on Computer Vision*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Navid Rajabi and Jana Kosecka. 2024. Q-groundcam: Quantifying grounding in vision language models via gradcam. *Preprint*, arXiv:2404.19128.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *Preprint*, arXiv:1506.01497.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Emmanuelle Salin, Badreddine Farah, S. Ayache, and Benoit Favre. 2022a. Are vision-language transformers learning multimodal representations? a probing perspective. In *AAAI Conference on Artificial Intelligence*.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022b. Are vision-language transformers learning multimodal representations? a probing perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11248–11257.

Santi Seguí, Oriol Pujol, and Jordi Vitrià. 2015. Learning to count with deep object features. *CoRR*, abs/1505.08082.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. DeAR: Debiasing Vision-Language Models with Additive Residuals . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6820–6829, Los Alamitos, CA, USA. IEEE Computer Society.

Ladan Shams and Aaron R. Seitz. 2008. Benefits of multisensory learning. *Trends in Cognitive Sciences*, 12(11):411–417.

Lloyd S. Shapley. 1953. 17. a value for n-person games.

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

Qingyi Si, Yuanxin Liu, Zheng Lin, Peng Fu, Yanan Cao, and Weiping Wang. 2023. Compressing and debiasing vision-language pre-trained models for visual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 513–529, Singapore. Association for Computational Linguistics.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. 2022. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171.

Fei Wang, Liang Ding, Jun Rao, Ye Liu, Li Shen, and Changxing Ding. 2024a. Can linguistic knowledge improve multimodal alignment in vision-language pretraining? *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(12).

Jin Wang, Shichao Dong, Yapeng Zhu, Kelu Yao, Weidong Zhao, Chao Li, and Ping Luo. 2024b. Diagnosing the compositional knowledge of vision language models from a game-theoretic view. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50332–50352. PMLR.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. Equivariant similarity for vision-language foundation models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11964–11974.

Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. 2022a. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812.

Yuxuan Wang, Difei Gao, Licheng Yu, Stan Weixian Lei, Matt Feiszli, and Mike Zheng Shou. 2022b. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022c. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France. European Language Resources Association.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256.

Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.

Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. 2021. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. In *Advances in Neural Information Processing Systems*, volume 34, pages 4514–4528. Curran Associates, Inc.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *Preprint*, arXiv:2306.13549.

Dani Yogatama, Lingpeng Kong, and Noah A. Smith. 2015. Bayesian optimization of text representations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2100–2105, Lisbon, Portugal. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. 2024. Investigating Compositional Challenges in Vision-Language Models for Visual Grounding . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14141–14151, Los Alamitos, CA, USA. IEEE Computer Society.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, MinQi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024a. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*.

Yi Zhang, Ce Zhang, Yushun Tang, and Zhihai He. 2024b. Cross-modal concept learning and inference for vision-language models. *Neurocomputing*, 583:127530.

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

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2023. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *Preprint*, arXiv:2207.00221.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2017. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *Preprint*, arXiv:2410.03659.

Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. 2022. Diagnosing vision-and-language navigation: What really matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5981–5993, Seattle, United States. Association for Computational Linguistics.

A Glossary

Modality Contribution The degree of each modality contributes towards the final prediction in downstream tasks. Modality contribution aims to quantify how important each modality is when generating output, and each modality is given an important score (Parcalabescu and Frank, 2023).

Cross-Modal Interaction How different modalities relate with each other and provide new information that unimodal alone cannot achieve (Liang et al., 2023b). For example, in VQA setting, given an image of an apple on a table and the text query *"Where is the apple?"*, it is impossible to provide correct answers using text or images only. Both inputs have to be utilized in order to generate the correct answer.

Dominant Modality The modality that weights more during inference (Frank et al., 2021; Liang et al., 2023a). Dominant modality could exist either naturally or due to unintended factors. Naturally, certain task setups are designed to rely more on one modality, where additional modalities provide supplementary information. Unintentionally, a model may learn to rely more heavily on one modality over the other(s) due to an imbalanced or coarse-grained pretraining dataset.

Unimodal Collapse A unimodal model achieves similar accuracy on a vision-language task compared to a multimodal model (Parcalabescu and Frank, 2023). This is an antonym for *dominant modality*. For instance, the input to a model consists of modality A and B, if modality A is the dominant modality, then modality B is "collapsed".

B Benchmark Datasets

FOIL IT! (Shekhar et al., 2017) is an extension of COCO dataset (Lin et al., 2014). It modifies the text description in each text-image pair to contain one mistake (so-called 'foil word'). Experimental results showed that VLMs perform poorly on caption classification, foil word detection, and foil word correction.

VALSE (Parcalabescu et al., 2022) contains six tasks: linguistic phenomena, including existence, plurality, counting, relations, actions, and coreference. Each instance is a correct caption, a foiled caption, and an image. The task is to ask a model to select the correct captions from foils. Experimental results showed that VLMs can identify the existence of an object, but fail to ground other linguistic phenomena.

Winoground (Thrush et al., 2022) is a small-scale visio-linguistic compositional reasoning dataset annotated by expert annotators with experience in vision and language research. In this dataset, two images and two text captions are considered as a pair. Both captions contain the exact same set of words, but in a different order. Diwan et al. (2022) showed that Winoground required commonsense reasoning and complex localization, which is beyond the scope of vision-linguistic compositional reasoning.

VSR (Visual Spatial Reasoning) (Liu et al., 2023a) contains over 10k text-image pairs with 66 types of spatial relations (e.g., under, facing). Both text and image are randomly sampled from MS COCO dataset. Each pair of images are then labeled by human annotators, such that the caption is correct for one image, and incorrect for another one, determined by spatial relations only.

ARO (Yuksekgonul et al., 2023) Attribution, Relation and Order Benchmark (ARO) is a fine-grained dataset for relation, attribution, and order understanding. The authors utilize Visual Genome (VG) and GQA for relations and attribution understanding probing tasks. They also utilize COCO Order (Lin et al., 2014) and Flicker30k Order (Young et al., 2014), by perturbating image captions to test VLMs' sensitivity to word order.

CREPE (Ma et al., 2023b) aims to test VLMs ability to generalize knowledge to unseen data (e.g., "red apple" in training, "green apple" in testing) and increasingly complex compositions (e.g., multiple attributes or relationships). It draws from existing datasets such as CC-12M, YFCC-15M, and LAION-400M. The authors filter and split the dataset into seen/unseen atoms, and increasingly complex scenes.

EQBEN (Wang et al., 2023) is a challenging VL compositionality dataset where it defines a stricter rule for "minimal semantic change". Specifically, it utilizes temporal frame changes in video dataset(e.g., Action Genome (Ji et al., 2019), GEBC (Wang et al., 2022b), and YouCook2 (Zhou et al., 2017)) to achieve minimal semantic difference between text and image pairs.

SVO-Probes (Hendricks and Nematzadeh, 2021b) tests VLMs verb understanding. The authors first created a large set of verb lists from Conceptual Captions dataset and generate negative samples by replacing the subject, verb, and object from the original caption. The images are collected from Google Image and verified via crowd-sourcing.

VL-Checklist (Zhao et al., 2023) uses four existing datasets: VG (Krishna et al., 2016), SWIG (Pratt et al., 2020), VAW (Pham et al., 2021), and HAKE (Li et al., 2019b) and transformed their original captions into incorrect captions. It aims to measure the ability of VLMs to detect incorrect object, attribute and relation.

MMVP (Tong et al., 2024) stands for Multimodal Visual Patterns, is a human-annotated benchmark dataset consisting of 9 visual patterns. Images are first collected by choosing samples that are contradicted in DINO and CLIP (i.e., high text-image similarity for one encoder but low in another). Human annotators then create captions and multiple-choice questions.

Predicate-Noun (Nikolaus et al., 2022) test VLMs' ability to understand relationships between a subject and its descriptor. Images from this dataset are collected from Open Images (Kuznetsova et al., 2018), where the authors manually verify examples and corresponding counterexamples, to ensure that counterexamples serve as strong distractors.

C Probing Studies

C.1 Linguistic Comprehension

Probing studies under linguistic comprehension focus on the ability of VLMs to comprehend textual input and extract meaningful patterns. Tasks in this category include negation, verb, and predicate-noun understanding. These tasks evaluate whether VLMs can handle syntactic and semantic nuances for language comprehension.

Verb understanding probes VLMs ability to comprehend actions or states described in textual input. For instance, given an image of a person running and a caption, "*The person is running*", the model should be able to match the alignment and able to identify mismatch with the caption says, "*The person is sitting*;;.

Predicate-noun understanding examines the model's ability to understand relationships between

Name	Paper	VQA	QA	Sent Anal.	Fusion	Retrieval	Vid. Reason.	Visual Ent.	Vis. Reason.	MM. Class.
					Modality Contribution					
MM-SHAP	(Parcalabescu and Frank, 2023)	✓								
Perceptual Score	(Gat et al., 2021)	✓					✓			
					Cross-Modal Interaction					
EMAP	(Hessel and Lee, 2020)	✓								✓
MultiViz	(Liang et al., 2023b)		✓		✓	✓				
DIME	(Lyu et al., 2022)	✓	✓							
IMF	(Xue et al., 2021)	✓						✓	✓	
M2Lens	(Wang et al., 2022a)			✓						

Table 1: List of metrics that evaluate the contribution of different modalities and cross-modal interaction and the task they have been evaluated on. Abbreviations: **VQA**: *Visual Question Answering*; **QA**: *Question Answering*; **Sent Anal.**: *Multimodal Sentiment Analysis*; **Fusion**: *Vision-Language Fusion*; **Retrieval**: *Vision-Language Retrieval*; **Vid. Reason.**: *Video Reasoning*; **Visual Ent.**: *Visual Entailment*; **Vis. Reason.**: *Visual Reasoning*; **MM. Class.**: *Multimodal Classification*

a subject (predicate) and its descriptor (noun). For example, given an image of a small cat and a caption, "*The small cat is on the mat*", the model should recognize the link between "small" and "cat".

Negation understanding tests VLMs' ability to understand negated statements. For an image of a red ball, the model should interpret the negation in the caption, "*The ball is not red*", and identify it as negative caption. Another example is, "*A beach with people*" and "*A beach without people*".

C.2 Visual Perception

Visual perception tasks examine the capacity of VLMs to interpret visual information. Studies in this category assess the model's ability to understand counting, attributes (e.g., identifying shapes, materials, colors, or sizes), and spatial reasoning. These tasks measure how well VLMs encode visual details and align them with corresponding textual descriptions, bridging vision and language.

Color understanding tasks test a model's ability in recognizing the color of a specific object in an image. For example, given an image of a green banana and query "*What is the color of the banana?*", VLMs that fail to perceive visual information will rely on textual bias, and answer "*Yellow*", given that most bananas are yellow color.

Existence tasks test the model's ability to detect the presence or absence in a scene. Given an image of a dog and a query, "*Is there a dog in this picture?*", the model should verify the dog's presence and provide "*Yes*" as a response.

Counting tasks require models to determine the number of objects in an image. For instance, given an image of three apples and a query, "*How many apples are in the image?*", the model should provide response, "*Three apples*". Incorrect response shows that the model fails to count from the input image.

C.3 VL Composition

VL compositionality tasks require VLMs to integrate visual and textual input to generate meaningful, cross-modal representations. It involves combining the semantics of linguistic comprehension with fine-grained visual features that require visual perception.

Attributes evaluates whether a model can identify characteristics related to an object, such as color, shape, or size. For example, given an image of a yellow triangle and a caption, "*The triangle is yellow*", the model should confirm the attribute. Conversely, it should flag a mismatch for a caption like "*The triangle is blue*".

Spatial relationships evaluates a model's understanding of object arrangements within the image. For example, given an image of a ball under a table and a caption, "*The ball is under the table*", the model should confirm the spatial relationship and identify mismatch like, "*The ball is on the table*".

Relations assess whether a model can capture interactions between multiple objects in an image. For instance, given an image of a god chasing a ball, the model should correctly verify the caption "*The dog is chasing the ball*" while rejecting "*The ball is chasing the dog*" as incorrect. Proper understanding of object relations is crucial for VLMs to generate accurate and contextually grounded descriptions.

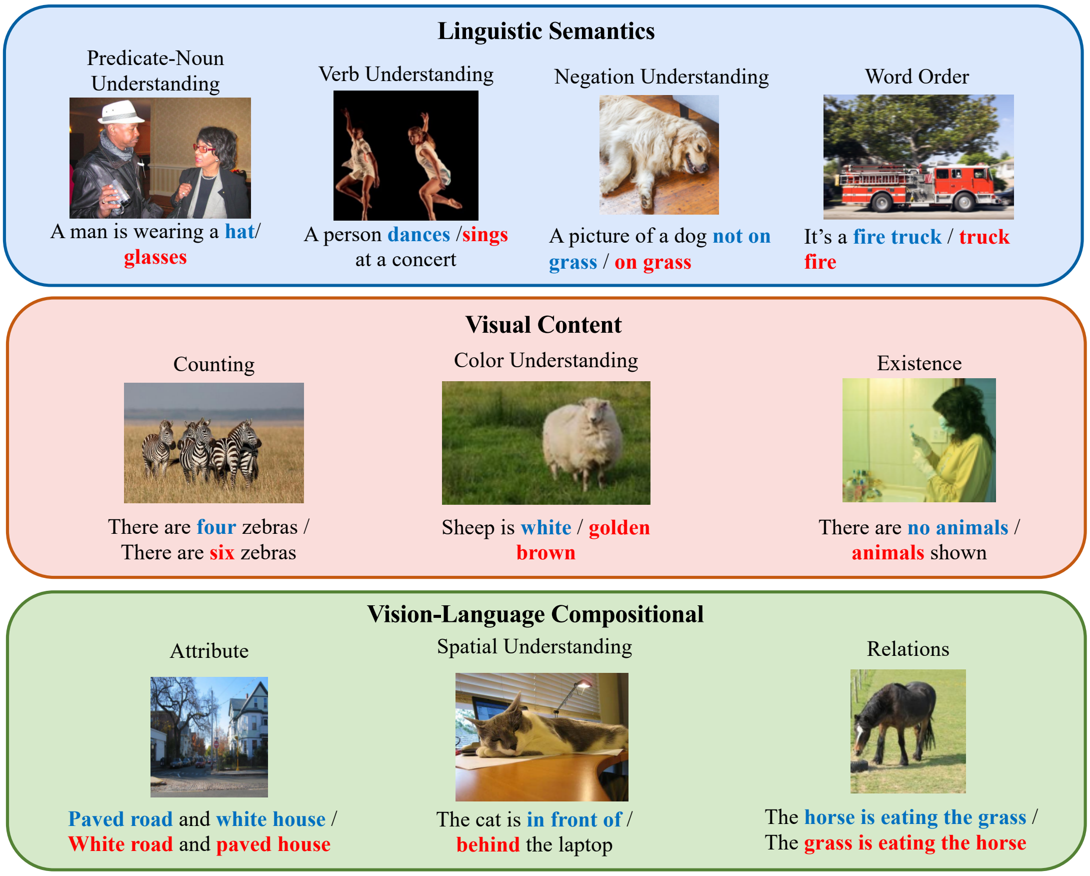

Figure 5: Example for each groups in linguistic semantics, visual content and vision-language compositional. Blue words indicate positive examples, while red words denote negative examples.

D Cross-Modal Interaction and Modality Contribution Metrics

Table 1 shows a list of metrics for modality contribution and cross-modal interaction and tasks that are evaluated in the original paper.

E Examples for Probing Tasks

Figure 5 shows example for each category described in Section 4.