Review Article

Large Vision-Language Model Alignment and Misalignment: A Survey Through the Lens of Explainability

Dong Shu[1], Haiyan Zhao[2], Jingyu Hu[3], Weiru Liu[3], Lu Cheng[4], Mengnan Du[2]

1. University of Northwestern, Saint Paul, United States; 2. New Jersey Institute of Technology, United States; 3. University of Bristol, United Kingdom; 4. University of Illinois at Chicago, United States

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in processing both visual and textual information. However, the critical challenge of alignment between visual and linguistic representations is not fully understood. This survey presents a comprehensive examination of alignment and misalignment in LVLMs through an explainability lens. We first examine the fundamentals of alignment, exploring its representational and behavioral aspects, training methodologies, and theoretical foundations. We then analyze misalignment phenomena across three semantic levels: object, attribute, and relational misalignment. Our investigation reveals that misalignment emerges from challenges at multiple levels: the data level, the model level, and the inference level. We provide a comprehensive review of existing mitigation strategies, categorizing them into parameter-frozen and parameter-tuning approaches. Finally, we outline promising future research directions, emphasizing the need for standardized evaluation protocols and in-depth explainability studies.

1. Introduction

Recent Large Vision-Language Models (LVLMs) have achieved significant progress in multimodal understanding. Models such as GPT-4V[1], Gemini[2], LLaVA[3], MiniGPT-4[4], and LLaMa 3.2[5] demonstrate unprecedented capabilities in tasks like image captioning and visual question answering, not only processing visual and textual information independently but also reasoning across these modalities. These advances are built upon two fundamental pillars: large language models (LLMs) and vision encoders. LLMs such as GPT-3.5[6], LLaMA[7], LLaMA 2[8], Vicuna[9], and

Qwen[10] revolutionized natural language processing, while vision encoders like CLIP[11] have transformed the ability to create aligned visual and textual representations, enabling effective vision-language understanding.

The key challenge in developing effective LVLMs lies in achieving proper alignment between visual and linguistic representations[12]. The predominant approach involves using representation alignment techniques, where visual features from an image encoder and textual representations from an LLM are mapped into a shared embedding space, typically matching the LLM's embedding dimensions[13][14][15]. Once both modalities are mapped into this shared space, alignment can be achieved through various training objectives and architectural designs that encourage the model to understand and reason about cross-modal relationships. This method has gained popularity due to its straightforward approach and generalizability across different model architectures.

However, the current understanding of alignment mechanisms remains limited. A critical challenge lies in misalignment phenomena, which manifest in various forms. For instance, when shown an image of a green apple, the model might fail to recognize the apple altogether (object misalignment), incorrectly describe it as red (attribute misalignment), or generate incorrect relationships like "the apple is floating in the air" when it's sitting on a table (relational misalignment). These misalignments lead to reliability issues[16][17][18], where models generate textual outputs that are inconsistent with the visual input. Understanding and addressing these misalignment issues is crucial for developing more reliable and trustworthy LVLMs, as they directly impact the models' ability to generate accurate and consistent multimodal outputs.

In this survey, we present a structured framework for understanding and addressing alignment challenges in LVLMs from an explainability perspective. We first examine the fundamentals of alignment, including its representational and behavioral aspects, training procedures, and theoretical foundations. We then analyze misalignment phenomena across three semantic levels: object, attribute, and relational misalignment. Our investigation reveals that misalignment stems from challenges at the data level (e.g., quality and balance issues), model level (e.g., architectural limitations and ability gaps), and inference level (e.g., task discrepancies). We review existing mitigation strategies and outline future research directions, emphasizing the need for standardized evaluation protocols and in-depth explainability studies.

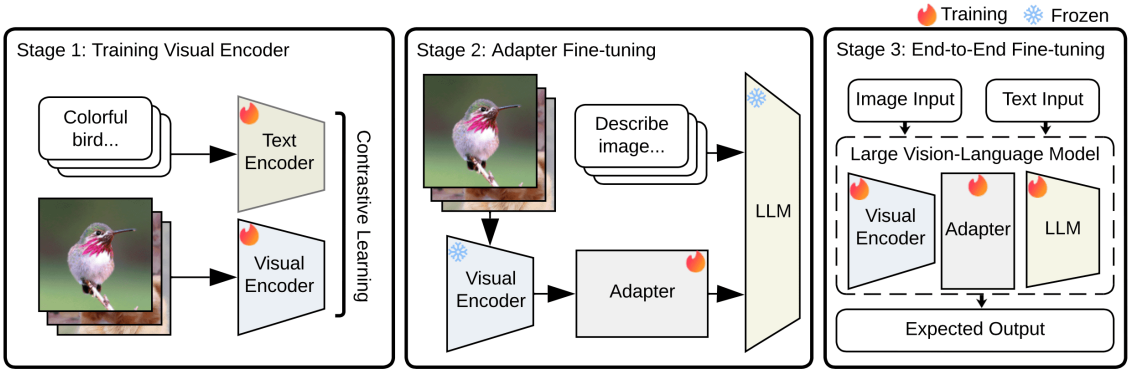

Figure 1. Overview of the three-stage LVLM training process, showing the progression from contrastive learning of visual-text encoders, through adapter fine-tuning with frozen components, to end-to-end model training.

2. Alignment of LVLMs

In this section, we examine alignment in LVLMs across four essential dimensions. First, we define the concept of alignment in LVLMs. Second, we detail the procedural stages through which alignment is achieved in practice. Third, we explore the theoretical foundations that make alignment possible between visual and linguistic modalities. Finally, we discuss methods for measuring and evaluating alignment in LVLMs.

2.1. What is Alignment?

In the context of LVLMs, let $\mathcal{X}$ be the image space and $\mathcal{T}$ be the text space. We define the alignment in two fundamental aspects: representational alignment and behavioral alignment.

- *Representational alignment* refers to the degree of correspondence between visual representations $v \in \mathcal{V}$ and textual representations $t \in \mathcal{T}$ within the model's internal embedding space $\mathcal{E}$. When well-aligned, the visual features extracted from an image and the textual embeddings of its corresponding description occupy nearby regions in the shared latent space, exhibiting high semantic similarity $d(v, t)$ where d is a similarity metric. This internal alignment enables the model to establish meaningful connections between visual and linguistic information at a fundamental level.

doi.org/10.32388/OOAM1Q

- *Behavioral alignment* refers to the model's ability to generate accurate, factual, and consistent textual responses $y \in \mathcal{Y}$ when processing image inputs $x \in \mathcal{X}$. A behaviorally aligned LVLM can reliably answer questions about visual content, provide precise descriptions, and perform reasoning tasks without introducing errors or hallucinations. This external manifestation ensures that the model's outputs faithfully reflect the actual content and relationships present in the images.

These two aspects of alignment are inherently connected. Strong representational alignment typically supports better behavioral alignment, as the model can more effectively leverage both visual and textual information to generate reliable outputs. Conversely, poor alignment in either aspect can lead to issues such as mismatched representations, inaccurate responses, or hallucinated content.

2.2. *How is Alignment Achieved?*

The development of alignment in LVLMs progresses through three major stages (see Figure 1), each is built upon its predecessor to achieve increasingly sophisticated cross-modal integration.

Stage 1: Training Visual Encoders. The foundation of LVLM alignment begins with training visual encoders through contrastive learning, exemplified by models like CLIP[11]. In this stage, the model learns to align visual and textual representations in a shared embedding space through a contrastive loss function. The process involves training on large-scale image-text pairs where matching pairs are pulled together in the embedding space while non-matching pairs are pushed apart. This leads to the development of robust visual representations that can meaningfully correspond to textual descriptions. Through this process, a visual encoder is created that can extract semantically meaningful features from images in a way that naturally aligns with language. This initial stage is crucial as it establishes the basic capability for cross-modal understanding, though the alignment is still relatively coarse-grained.

Stage 2: Adapter Fine-tuning. The second stage involves fine-tuning an adapter module that bridges the pre-trained visual encoder with the language model. This stage introduces lightweight adapter architectures, which typically consist of simple components such as linear layers, MLPs, or cross-attention layers that learn to translate between visual and language model embedding spaces. For example, cross-attention layers can feed image encoder representations into the language model, enabling the model to attend to relevant visual features when generating text[19]. A key characteristic of this approach is the preservation of the original capabilities of both the visual encoder and language

model while learning to interface between them. During adapter training, while the visual encoder parameters may be updated, the language model parameters often remain frozen to maintain their original text capabilities. This intermediate stage is essential for establishing effective connections between modalities while preserving the specialized capabilities of each component.

Stage 3: End-to-End Fine-tuning. The final stage involves comprehensive fine-tuning of the entire system, including the visual encoder, adapter, and LLM components together. This comprehensive approach allows for deeper integration and more sophisticated alignment between all components. It enables the model to learn task-specific optimizations that require coordinated adjustments across all modules. Through this process, the model develops more advanced cross-modal understanding capabilities and facilitates the emergence of emergent behaviors that arise from the deep integration of visual and linguistic processing. This stage often results in the highest performance but requires careful balancing to avoid catastrophic forgetting or degradation of pre-existing capabilities.

2.3. Why is Alignment Possible?

Having established what alignment means and how it is implemented in LVLMs, a fundamental question arises: why is such alignment between vision and language modalities possible in the first place? The possibility of alignment between these modalities can be understood from both theoretical and algorithmic perspectives.

Theoretical Perspective. From a theoretical standpoint, visual and textual data are different projections of the same underlying reality. As Huh et al. argue in their Platonic Representation Hypothesis[20], all modalities are measurements of a real world that generates our observations. When humans create images or write text, they are encoding information about this same reality, albeit through different measurement processes. Although these modalities appear distinct on the surface, they fundamentally capture overlapping semantic information about the same world state. This shared origin in physical reality, combined with the fact that humans generate both types of data to describe their observations of the world, provides the theoretical foundation for why these modalities can be meaningfully aligned in a common representation space.

Algorithmic Perspective. From an algorithmic perspective, although visual encoders and language models are initially trained separately on different modality-specific data, their learned representations inherently capture some similar semantic structures due to their training on human-generated data. Recent research has shown that these inherent similarities exist even before explicit

alignment training[21][22][23]. This natural compatibility serves as a starting point for more sophisticated alignment. The staged training process described in Section 2.2 then is built upon this inherent compatibility through systematic refinement: first using contrastive learning to organize embeddings in the shared latent space, then employing adapter fine-tuning to bridge between modalities while preserving their specialized capabilities, and finally conducting end-to-end training to enable deep integration across all components. Through this systematic combination of training stages and optimization objectives, the model gradually develops a robust alignment between the two modalities.

2.4. How to Measure Alignment?

This section examines approaches for quantifying the effectiveness of alignment in LVLMs. These measurement approaches naturally align with our earlier definition in Section 2.1 of representation alignment and behavioral alignment, and can be organized along these two fundamental levels.

Representation Level. At the representation level, alignment can be directly measured between visual and textual representations within the LVLM's embedding space by assessing how similarly the visual and textual modalities encode and relate to the same concepts or data points. The simplest approach is to compute the cosine similarity between the embeddings of visual and textual data. High alignment corresponds to scores close to 1, while low alignment corresponds to scores closer to 0[15]. More sophisticated metrics have been developed to assess alignment between the two representation spaces. For instance, the mutual nearest-neighbor metric quantifies alignment by evaluating the consistency of nearest neighbors across modalities[20]. Another approach is kernel alignment, which evaluates the similarity of pairwise relationships within each modality's embedding space, providing a holistic view of the alignment structure[21].

Behavioral Level. The behavioral level measures alignment through the model's performance on various downstream tasks and benchmarks, using both direct comparisons and automated evaluation systems. The strength of alignment directly impacts the LVLM's performance, as better alignment typically leads to improved task outcomes. These measurements generally involve comparing the model's outputs against ground truth labels, either through direct comparison or using evaluation models to simulate human judgment. Numerous benchmarks have been developed to assess LVLM alignment across a range of tasks, from coarse-grained evaluations (e.g., object existence) to fine-grained assessments (e.g., color, count, spatial relations). Examples of such benchmarks include

POPE[24], CHAIR[25], MME[26], MMHal-Bench[27], and LLaVa-Bench[3]. In addition to traditional benchmarks, advanced evaluation models like GAVIR[28], CCEval[29] and HaELM[30] provide sophisticated assessments by considering context and evaluating responses comprehensively, similar to human evaluators. The flexibility and diversity of evaluation models enable thorough measurement capabilities needed for open-ended questions.

3. Misalignment of LVLMs

After introducing the alignment of LVLMs, we now examine a critical challenge facing these models: their tendency to generate outputs that diverge from the visual input. Despite significant advances in alignment techniques, LVLMs still frequently exhibit misalignment between their visual and textual inputs. In this section, we provide a comprehensive analysis of misalignment phenomena in LVLMs, beginning with a definition and taxonomy of different types of misalignment (see Figure 2), followed by an examination of their underlying causes.

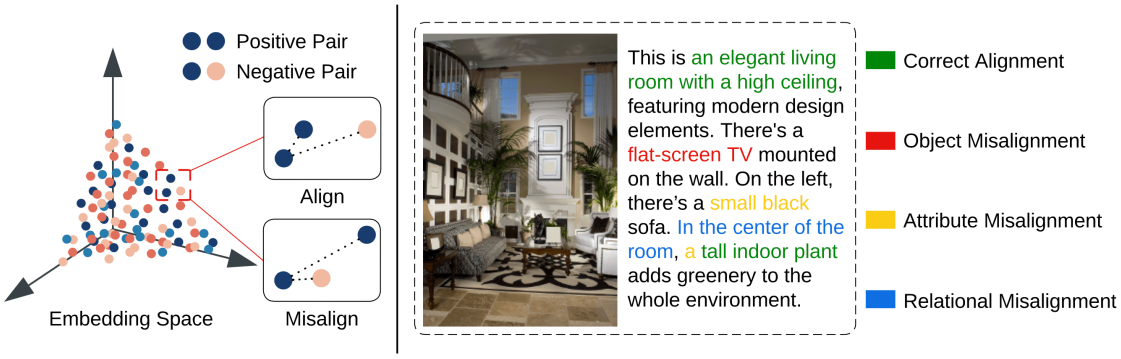

Figure 2. Illustration of representation-level and behavior-level alignment and misalignment in LVLMs. The **left** side shows **representation-level** phenomena in embedding space, where aligned visual-text pairs cluster together (positive pairs) while misaligned pairs are separated (negative pairs). The **right** side demonstrates **behavior-level** alignment and misalignment through a room description example, showing the spectrum from correct alignment (green) to various types of semantic misalignment: object misalignment (red), attribute misalignment (yellow), and relational misalignment (blue). These two levels are inherently connected, as the quality of representation alignment in the embedding space influences the model's ability to generate semantically aligned outputs.

3.1. Definition of Misalignment

Misalignment in LVLMs occurs when the model's output semantically diverges from the visual content it is meant to describe. These discrepancies show in several key phenomena, impacting the overall performance of these models. In this paper, we categorize behavior-level misalignment phenomena in LVLM into three semantic levels $\mathcal{S} = \{s_o, s_a, s_r\}$: *object misalignment* (s_o), *attribute misalignment* (s_a), and *relation misalignment* (s_r). Rather than using the term 'hallucination' commonly found in the literature[12], we adopt the term 'misalignment' to better characterize how these discrepancies emerge between visual and language representations.

- *Object Misalignment* (s_o): This is one of the most widely recognized forms of misalignment[12][24][30]. It occurs when the model generates descriptions containing objects O' that differ from the actual objects O in the image, where $O' \nsubseteq O$. This represents the most coarse-grained level of misalignment, as it simply refers whether an object exists in the image or not. Due to its coarse-grained nature, object misalignment is relatively straightforward to detect and mitigate.
- *Attribute Misalignment* (s_a): At a finer level, we identify attribute misalignment[31]. This occurs when for an object $o \in O$, the model correctly identifies the object but generates incorrect attributes $A' \neq A$, where A represents the true attributes of o. Attribute misalignment typically involves adjectives or adverbs that describe properties of objects inaccurately. For example, when input an image of a green apple, the model might incorrectly describe the color of an apple as 'red' instead of 'green'.
- *Relation Misalignment* (s_r): This category involves the generation of incorrect or non-existent relationships R' between objects in an image[32], where R' differs from the true relationships R. This misalignment manifests in two primary ways: spatial relationship errors and action relationship errors. In spatial relationships, the model might incorrectly describe the relative positions of objects, such as saying 'next to' when the correct relation is 'on top of', or 'inside' when objects are merely 'near' each other. In action relationships, the model might generate semantically impossible interactions between objects, such as 'he is walking a car' instead of 'he is driving a car', or 'the cat is reading a book' instead of 'the cat is sitting on a book'.

3.2. Reasons of Misalignment

Having identified the three semantic levels of misalignment phenomena, we now analyze their root causes across three fundamental levels: Dataset, Model, and Inference. The Dataset level examines how training data characteristics influence misalignment during learning. The Model level investigates how architectural decisions and training procedures affect alignment between modalities. The Inference level explores how the generation process can introduce misalignment even with well-aligned underlying representations.

3.2.1. Dataset Level

Data quality and distribution patterns play crucial roles in contributing to misalignment between visual and language representations in LVLMs. Several key dataset factors can impede the model's ability to form accurate associations between visual inputs and textual descriptions, affecting both training effectiveness and inference performance.

- *Data imperfections*: This includes blurry images, vague or inaccurate captions, and mismatched image-caption pairs, which introduce significant challenges during training[33][34]. These quality issues manifest in various forms: images may suffer from poor resolution, inappropriate cropping, or visual artifacts; captions might contain grammatical errors, ambiguous descriptions, or factually incorrect information; and in some cases, the captions may describe content entirely unrelated to their paired images. These low-quality data points can distort the model's ability to form precise mappings between modalities, leading to outputs that fail to accurately reflect the input image and potentially establishing incorrect associations that persist through the training process.
- *Data Imbalance*: When certain classes or types of data are disproportionately represented, it skews the model's training process[28][35]. For example, visual question-answering datasets often overrepresent positive answers, subtly training the model to favor these outcomes while underperforming on underrepresented negative answers.
- *Data Inconsistency*: Inconsistencies exacerbate misalignment by introducing contradictory outputs across different tasks for the same image. For instance, an image captioning task might describe an image as depicting 'a tiger eating a chicken,' yet in a visual question-answering task for the same

image, the answer to 'what is the tiger eating?' might label the prey as 'a duck'[36]. Such contradictions disrupt the model's ability to generate coherent and consistent outputs across tasks.

- *Data False Negative*: False negatives in the dataset further complicate alignment, as negative image-text pairs, though not perfectly matching, share overlapping components[37][38]. During training, embeddings of positive pairs are drawn closer together, while those of negative pairs are pushed apart. This binary method can suppress latent similarities within false negatives, reducing the model's capacity to generalize and effectively align diverse modalities.
- *Data Polysemy*: The inherent polysemy within datasets introduces additional complexity. Polysemy enriches data diversity by allowing a single word or image to convey multiple meanings depending on context, but this ambiguity also amplifies the risk of misalignment[39][40]. For example, an image caption of 'the bat hit the ball' could refer to the animal or the baseball bat. This variability challenges the model to establish consistent mappings between modalities.

3.2.2. Model Level

Beyond data-level issues, the architectural design and training methodology of LVLMs significantly influence model alignment.

- *Separate Training*: Prior to being integrated as a single LVLM, the visual encoder and the LLM are typically pre-trained separately on distinct single-modality datasets. While this approach offers advantages such as efficiency and modularity, it leads to each model developing its own biased representations and understanding of the world, shaped by its respective single-modality data[41].
- *Ability Gap*: This independent pretraining process also creates an ability gap between the visual encoder and the LLM[42], where the LLM often demonstrates significantly greater capability than the visual encoder. Consequently, the LVLM tends to rely excessively on the LLM for predictions, resulting in imbalanced attention between visual and textual information[43][44][45].
- *Pretrain-finetuning Knowledge Gap*: After integrating the visual encoder and LLM into a unified LVLM, fine-tuning is typically performed to further enhance alignment and adapt the model to specific downstream tasks. However, this fine-tuning phase can introduce a pretraining-finetuning knowledge gap or conflict, where the general knowledge acquired during pretraining may clash with the specific requirements of the fine-tuning task[17]. Such conflicts can lead to knowledge forgetting, where the LVLM loses previously learned information while adapting to the new task[46][47]. Although knowledge forgetting might appear insignificant, it can have cascading

effects. Each unit of knowledge in the model's embedding space is interconnected with lots of semantic relationships. Forgetting even a single piece of knowledge can disrupt these relational connections, undermining the integrity of the embedding space. This disruption causes a broader misalignment within the LVLM.

- *Knowledge Conflict*: A significant challenge arises from knowledge conflicts between the visual encoder and language model components of LVLMs. These conflicts emerge when the visual encoder's direct perception of image content contradicts the prior knowledge embedded in the LLM's parameters during pre-training[48][49]. For example, when an image contains a green tomato, the visual encoder accurately detects its color, but the LLM may resist this information since it has been predominantly trained on texts describing ripe, red tomatoes. This misalignment between observed visual evidence and learned textual priors can manifest in various ways: the model might incorrectly describe the tomato as red despite clear visual evidence, generate hesitant or self-contradicting descriptions, or attempt to rationalize the discrepancy by making unwarranted assumptions about the tomato's ripeness stage.

3.2.3. Inference Level

Misalignment can also occur during the inference stage due to *task discrepancy*. This discrepancy fundamentally represents an out-of-distribution (OOD) generalization problem, as users often pose questions or request tasks that deviate from the distribution of examples seen during training. Even when a LVLM has been trained on a large and diverse dataset, it may encounter novel combinations of visual and textual elements or be asked to perform tasks in ways that differ subtly but significantly from its training examples. This OOD challenge manifests in several ways. First, the training data used for pre-training or fine-tuning the model may not fully align with the specific tasks it is later expected to perform[16]. For example, a model trained primarily on image captioning data might struggle when asked to answer specific questions about spatial relationships or perform detailed visual reasoning tasks. Second, users may phrase requests in ways that differ from the instruction patterns seen during training, leading to potential misinterpretation of the task requirements. Third, the visual inputs during inference may contain novel object configurations or scene compositions not well-represented in the training data. These distribution shifts can create misalignment in LVLMs as the model struggles to adapt to new and distinct tasks that require different interpretations of visual and textual information.

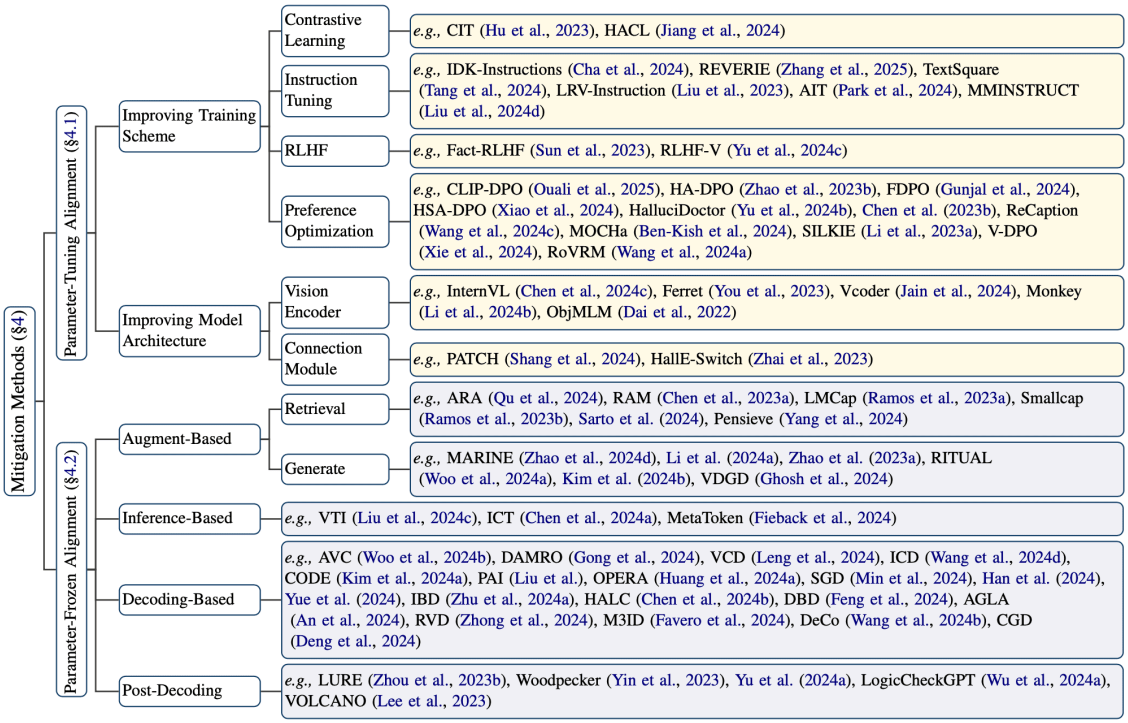

Figure 3. Taxonomy of Misalignment Mitigation Methods for LVLMs, including *Parameter-Tuning Alignment* and *Parameter-Frozen Alignment*. Refs: [27][28][29][31][33][35][44][45][49][50][51][52][53][54][55][56] [57][58][59][60][61][62][63][64][65][66][67][68][69][70][71][72][73][74][75][76][77][78][79][80][81][82][83][84][85][86] [87][88][89][90][91][92][93][94][95][96][97][98][99][100][101][102][103][104][105][106]

4. Mitigation Methods

Building upon our analysis of misalignment causes in LVLMs, we now examine strategies for mitigating these challenges (see Figure 3). These mitigation approaches can be categorized into two groups: parameter-tuning alignment methods and parameter-frozen alignment methods. Parameter-tuning alignment involves modifying specific components within the LVLM architecture to reduce misalignment through targeted parameter updates. In contrast, parameter-frozen alignment methods address misalignment while maintaining the LVLM's original parameters unchanged, offering solutions that preserve the model's structure while improving its cross-modal alignment capabilities.

4.1. Parameter Tuning Alignment

Parameter-tuning alignment focuses on mitigating misalignment by refining the training scheme or enhancing the architecture itself.

Improving Training Scheme. Parameter-tuning methods that improve the training scheme often address misalignment broadly as a data-level issue or as a general visual-textual misalignment[33][50]. This understanding leads to a straightforward objective, which is reducing the modality gap between visual and textual representations. This can often achieved by improving the dataset quality or optimizing training techniques. One common approach is contrastive learning, exemplified by methods such as CIT[35] and HACL[50]. These techniques involve using a third model to generate positive and negative data pairs. The LVLM is then trained to bring the representations of positive pairs closer together while pushing negative pairs apart in the embedding space. Another widely adopted strategy is instruction tuning, as seen in LRV-Instruction[28] and TextSquare[53]. Similarly, these approaches rely on a third model to generate instructional data, which is subsequently used to train the LVLM effectively. However, these approaches often lack robust quality assurance mechanisms to verify the accuracy or relevance of the generated data, introducing potential risks. Alternatively, Reinforcement Learning from Human Feedback (RLHF) employs human feedback to train a reward model, ensuring that the generated data aligns with human preferences[27][56]. While RLHF guarantees high-quality training data, it comes at a significant cost. To address this, some methods leverage preference optimization, wherein multiple responses are generated for the same input image, ranked or scored by a third model, and categorized into positive and negative pairs[33][57][58]. The model is then fine-tuned on this curated dataset. Although these methods can significantly improve the model, they are often constrained by either high resource requirements (as in RLHF) or the uncertain quality of generated data (as in contrastive learning and instruction tuning) or rerank model (as in preference optimization). This highlights the ongoing need for large, diverse, and high-quality datasets to effectively address data-level misalignment.

Improving Model Architecture. Methods that improve the model architecture often involve a deep understanding of the root causes of misalignment, allowing researchers to pinpoint deficiencies within specific components of the LVLM. Typical LVLM architectures consist of three main components: the visual encoder, the adapter module, and the LLM[12][107]. Most architecture-focused approaches concentrate on enhancing the visual encoder or the adapter module, with relatively few

addressing improvements to the LLM itself. This aligns with our earlier model-level claim of the model ability gap, where the LLM often outperforms the visual encoder. Blindly enhancing the LLM could exacerbate this gap, potentially worsening the misalignment issue. To reduce this ability gap, some studies scale up the visual encoder by increasing its parameter size[67]. Others introduce additional components to the visual encoder to improve its capabilities without necessarily scaling up its size[68][69][70]. In addition to the visual encoder, many methods focus on improving the adapter module, which serves as the critical bridge between the visual and textual modalities. Enhancements to the adapter module often involve adding intermediary layers or mechanisms to better align the visual encoder's outputs with the LLM's input requirements. For example, PATCH[31] employs trainable virtual tokens to enhance the projection layer, improving cross-modal alignment. Similarly, HallE-Switch[29] introduces a dynamic mechanism that adjusts the flow of information between the visual encoder and the LLM based on input complexity. By addressing these architectural components, parameter-tuning methods aim to reduce the modality gap and improve the alignment between visual and textual representations, ultimately enhancing the LVLM's performance across tasks.

4.2. Parameter Frozen Alignment

Parameter-frozen alignment methods have gained increasing popularity due to their significant practical advantages. These training-free approaches are highly modular and easy to implement, allowing them to be readily integrated into existing systems without requiring costly retraining or fine-tuning processes. This makes them particularly attractive for real-world applications where computational resources may be limited. We categorize these parameter-frozen methods into four types based on where they intervene in the LVLM processing pipeline: Augment-based mitigation, augmenting the LVLM by incorporating external knowledge; inference-based mitigation, operating in the model's latent space during intermediate processing; decoding-based mitigation, which guides the text generation process; and post-decoding mitigation, which refines the final outputs.

Augment-based Methods. As analyzed in Section 3, insufficient input of image information is one of the primary causes of misalignment, leading to poor visual understanding. To address this, retrieval-augmented generation (RAG) methods have been adapted to dynamically integrate external knowledge into LVLMs through retrieved results[80][72][73][74][75][76]. By reranking the similarity of image-text pairs, RAG approaches provide more visual context and guidance to the model. Similarly, other methods rely on generating approach to enrich the input with additional information. For

instance,[77],[78], and[79] propose integrating an auxiliary model to generate relevant information based on the image. Alternatively, methods such as RITUAL[81] bypass the need for external models. It enhances the model's exposure to diverse visual contexts by applying random transformations to input images. Additionally, approaches like[82][49] employ self-generated textual descriptions appended to the input prompt, ensuring the model has sufficient knowledge to answer questions accurately without generating plausible but incorrect responses.

Inference-based Methods. Some methods operate in the model's latent space during the inference process, prior to decoding, by intervening in both visual and textual representations to improve alignment. For instance, Visual and Textual Intervention (VTI)[83] pre-computes intervention directions using a small set of examples and applies them during inference to enhance feature stability and vision-text alignment, without requiring additional training. Similarly, Image-Object Cross-Level Trusted Intervention (ICT)[84] introduces a lightweight mechanism that intervenes in the model's attention at both image and object levels, applying targeted activation shifts to selected attention heads. Since they operate directly on the model's internal representations, they can make precise adjustments to improve alignment without disrupting the model's broader language understanding capabilities. This makes inference-based methods particularly effective at reducing misalignment while preserving the model's ability to generate fluent and contextually appropriate responses.

Decoding-based Methods. Another widely used approach for mitigating misalignment involves modifying decoding process. These methods often target issues of imbalanced attention. However, the imbalance attention between what still remain debated. Some researchers argue that the model over-focuses on irrelevant image tokens, such as background elements or unimportant details[45][86]. However, the prevailing view is that the model prioritizes textual tokens over visual ones, neglecting critical visual information[87][88][89][90]. Despite these differences in interpretation, most decoding-based methods employ similar contrastive decoding techniques to rebalance attention between modalities, typically by reducing attention to textual tokens while enhancing focus on visual tokens. This approach, however, contrasts with inference-based methods, which avoid reducing attention to textual information and instead preserve the model's overall language understanding. Another interesting observation is that, while decoding-based methods typically lead to similar approaches, in some cases, they can result in fundamentally divergent strategies. For instance, OPERA[91] hypothesizes that the model over-relies on summary tokens, instead of focusing visual

tokens. However, text summarization is SGD's solution[44] to mitigate misalignment. It uses summarization to shorten textual context and helps model shift focus toward visual information. This divergence underscores how subtle differences in understanding the root causes of misalignment can result in contradicted methodologies.

Post-decoding Methods. Lastly, post-decoding approaches present a broader range of hypotheses about the causes of misalignment, tackling issues ranging from data-level biases to model-level deficiencies. Methods such as LURE[102] and Woodpecker[103] exemplify this category. LURE focuses on addressing object hallucinations by revising the generated text, identifying hallucinatory content, and reconstructing less biased outputs. Woodpecker employs a five stages validation mechanism to extract and correct inconsistencies in the generated response. Despite their specific details, these methods converge on a shared strategy, which involves modifying the model's outputs after decoding without altering the its parameters or architecture, making them easily adaptable to various LVLMs. This flexibility lies in their goal-oriented nature, as they directly target specific misalignment phenomena. However, this goal-oriented focus introduces a significant limitation. While it can enhance output quality, the underlying model deficiencies are left unchanged, restricting generalization and limiting performance on tasks beyond post-decoding corrections.

5. Future Research Directions

In this section, we discuss several important directions for future research in understanding and improving alignment in LVLMs.

5.1. Standardized Benchmarks

The current evaluation of misalignment in LVLMs suffers from a critical limitation, i.e., the lack of standardized, comprehensive benchmarks that can systematically assess different types of misalignment across models. While existing benchmarks have made important contributions, they typically focus on specific aspects of misalignment in isolation. For instance, POPE[24] primarily evaluates object hallucination, while other benchmarks concentrate on particular relationship errors or attribute inconsistencies. What is urgently needed is a unified evaluation framework that can systematically assess misalignment across all semantic levels, from object-level (e.g., describing a non-existent dog in an image) to attribute-level (e.g., color, size, texture errors) and relation-level misalignment (e.g., spatial relationship errors). Such a comprehensive benchmark would enable direct

comparisons between different LVLM architectures and alignment techniques using standardized metrics, evaluate both representational alignment and behavioral alignment, and assess how misalignment manifests across different types of tasks. The benchmark should also consider both the frequency and severity of different types of misalignment, rather than treating all misalignments as equally problematic. The development of such standardized benchmarks would represent a significant step forward in our understanding of misalignment in LVLMs and accelerate progress toward more reliable and trustworthy vision-language systems.

5.2. Explainability based Diagnose

To better understand and address alignment issues in LVLMs, future research should leverage advanced explainability techniques that can reveal the internal mechanisms of these models. There are two critical categories of explainability approaches that warrant investigation: (1) internal knowledge decoding and (2) attribution methods.

The first category of explainability approaches centers on internal knowledge decoding and understanding how information is processed within LVLMs[108][109]. Mechanistic interpretability approaches could help identify specific components and circuits within LVLMs that are responsible for cross-modal alignment, providing insights into how visual and language representations are integrated and processed. Similarly, probing techniques can analyze the emergence and evolution of aligned representations across different layers and attention heads, helping researchers understand where and how misalignment occurs within the model architecture[110]. This detailed understanding of the internal working mechanisms would not only advance theoretical knowledge but also guide the development of more effective alignment techniques.

The second critical category focuses on attribution methods that can determine the relative influence of different information sources on model outputs. LVLMs have three primary information sources for generating outputs: user text prompts, input images, and knowledge stored within pre-trained LLMs. Future research needs to develop sophisticated attribution algorithms that can determine whether a model's output primarily depends on the input text prompt, derives from the visual information in the image, or relies on the LLM's internal knowledge. This detailed attribution analysis would help identify when and why misalignment occurs, such as cases where the model inappropriately relies on LLM knowledge rather than visual evidence, or when it fails to properly integrate information from

multiple sources. Such insights would be useful for designing targeted mitigation strategies that address specific types of misalignment and improve the overall reliability of LVLMs.

5.3. Architectural Innovations

Current LVLM architectures face fundamental challenges, including significant ability gaps between visual encoders and LLMs, persistent attention imbalances between modalities, and knowledge conflicts between visual and linguistic representations. While most existing solutions focus on improving training procedures or adding post-processing steps, future research should focus on architectural innovations that address these structural limitations. This could include developing novel integration mechanisms that better balance the capabilities of visual and language components, and dynamic architectures that can adaptively adjust their attention mechanisms to maintain equilibrium between modalities. The field would benefit from multi-stage processing architectures that explicitly manage knowledge conflicts through specialized components for different levels of semantic understanding. Additionally, new transformer architectures specifically designed for vision-language tasks, rather than adapted from unimodal architectures, could help bridge the ability gap between visual and linguistic processing.

6. Conclusions

In this paper, we present a systematic survey of alignment and misalignment in LVLMs through an explainability lens. Our investigation demonstrates that achieving proper alignment involves complex interactions between data quality, model architecture, and inference procedures. We have developed a categorization of misalignment into object, attribute, and relational levels, offering a clear framework for understanding these challenges and developing targeted solutions. The examination of current mitigation strategies reveals a spectrum of approaches, from computationally intensive parameter-tuning methods to more practical parameter-frozen solutions, each with distinct trade-offs between effectiveness and implementation feasibility. Lastly, we have identified several key directions for future research, which will be essential for creating more reliable and capable vision-language systems that maintain robust alignment while serving diverse real-world applications.

References

1. ^OpenAI. "ChatGPT can now see, hear, and speak." https://openai.com/index/chatgpt-can-now-see-hear-and-speak/. Accessed: September 25, 2023.
2. ^Gemini Team, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805. 2023.
3. a, bLiu H, Li C, Wu Q, Lee YJ (2024). "Visual instruction tuning". Advances in Neural Information Processing Systems. 36.
4. ^Zhu D, Chen J, Shen X, Li X, Elhoseiny M (2023). "Minigpt-4: Enhancing vision-language understanding with advanced large language models". arXiv preprint arXiv:2304.10592.
5. ^Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024.
6. ^Brown TB (2020). "Language models are few-shot learners". arXiv preprint arXiv:2005.14165.
7. ^Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al. (2023). "Llama: Open and efficient foundation language models". arXiv preprint arXiv:2302.13971.
8. ^Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. 2023.
9. ^Chiang WL, Li Z, Lin Z, Sheng Y, Wu Z, Zhang H, Zheng L, Zhuang S, Zhuang Y, Gonzalez JE, et al. (2023). "Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality". See https://vicuna. lmsys. org (accessed 14 April 2023). 2 (3): 6.
10. ^Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, Fan Y, Ge W, Han Y, Huang F, et al. 2023. "Qwen technical report". arXiv preprint arXiv:2309.16609.
11. a, bRadford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021. p. 8748–8763.
12. a, b, c, dLiu H, Xue W, Chen Y, Chen D, Zhao X, Wang K, Hou L, Li R, Peng W (2024). "A survey on hallucination in large vision-language models". arXiv preprint arXiv:2402.00253.
13. ^Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le Q, Sung YH, Li Z, Duerig T (2021). "Scaling up visual and vision-language representation learning with noisy text supervision". In: International conferen

ce on machine learning. PMLR. pp. 4904–4916.

14. ^Yang J, Duan J, Tran S, Xu Y, Chanda S, Chen L, Zeng B, Chilimbi T, Huang J (2022). "Vision-language pre-training with triple contrastive learning". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15671--15680.
15. a, bShu D, Duan B, Guo K, Zhou K, Tang J, Du M (2024). "Exploring the Alignment Landscape: LLMs and Geometric Deep Models in Protein Representation". arXiv preprint arXiv:2411.05316.
16. a, bZhang Y, Li J, Liu L, Qiang W (2024). "Rethinking Misalignment in Vision-Language Model Adaptation from a Causal Perspective". arXiv preprint arXiv:2410.12816.
17. a, bZhou Y, Cui C, Rafailov R, Finn C, Yao H (2024). "Aligning modalities in vision large language models via preference fine-tuning". arXiv preprint arXiv:2402.11411.
18. ^Zhao T, Zhang L, Ma Y, Cheng L. A survey on safe multi-modal learning systems. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024. p. 6655-6665.
19. ^Meta AI (2024). "Llama 3.2: Revolutionizing edge AI and vision with open, customizable models". Meta AI Blog. Retrieved December 20, 2024. Available from: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.
20. a, bHuh M, Cheung B, Wang T, Isola P (2024). "The platonic representation hypothesis". The International Conference on Machine Learning (ICML).
21. a, bManiparambil M, Akshulakov R, Djilali YA, El Amine Seddik M, Narayan S, Mangalam K, O'Connor NE (2024). "Do Vision and Language Encoders Represent the World Similarly?" In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14334–14343.
22. ^Sharma P, Shaham TR, Baradad M, Fu S, Rodriguez-Munoz A, Duggal S, Isola P, Torralba A (2024). "A vision check-up for language models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14410–14419.
23. ^Neo C, Ong L, Torr P, Geva M, Krueger D, Barez F (2024). "Towards Interpreting Visual Information Processing in Vision-Language Models". arXiv preprint arXiv:2410.07149.
24. a, b, cLi Y, Du Y, Zhou K, Wang J, Zhao WX, Wen JR (2023). "Evaluating object hallucination in large vision-language models". arXiv preprint arXiv:2305.10355.
25. ^Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K (2018). "Object hallucination in image captioning". arXiv preprint arXiv:1809.02156.
26. ^Fu C, Chen P, Shen Y, Qin Y, Zhang M, Lin X, Yang J, Zheng X, Li K, Sun X, et al. (2023). "MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models". arXiv preprint arXiv:2306.13

394.

27. a, b, cSun Z, Shen S, Cao S, Liu H, Li C, Shen Y, Gan C, Gui L, Wang Y, Yang Y, et al. (2023). "Aligning large multimodal models with factually augmented rlhf". arXiv preprint arXiv:2309.14525.
28. a, b, c, dLiu F, Lin K, Li L, Wang J, Yacoob Y, Wang L (2023). "Mitigating hallucination in large multi-modal models via robust instruction tuning". In: The Twelfth International Conference on Learning Representations.
29. a, b, cZhai B, Yang S, Zhao X, Xu C, Shen S, Zhao D, Keutzer K, Li M, Yan T, Fan X (2023). "Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption". arXiv preprint arXiv:2310.01779.
30. a, bWang J, Zhou Y, Xu G, Shi P, Zhao C, Xu H, Ye Q, Yan M, Zhang J, Zhu J, et al. (2023). "Evaluation and analysis of hallucination in large vision-language models". arXiv preprint arXiv:2308.15126.
31. a, b, cShang Y, Zeng X, Zhu Y, Yang X, Fang Z, Zhang J, Chen J, Liu Z, Tian Y (2024). "From Pixels to Tokens: Revisiting Object Hallucinations in Large Vision-Language Models". arXiv preprint arXiv:2410.06795.
32. ^Wu M, Ji J, Huang O, Li J, Wu Y, Sun X, Ji R (2024). "Evaluating and analyzing relationship hallucinations in large vision-language models". arXiv preprint arXiv:2406.16449.
33. a, b, c, dOuali Y, Bulat A, Martinez B, Tzimiropoulos G. 2025. "CLIP-DPO: Vision-Language Models as a Source of Preference for Fixing Hallucinations in LVLMs." In European Conference on Computer Vision, pages 395–413. Springer.
34. ^Shi Z, Wang Z, Fan H, Zhang Z, Li L, Zhang Y, Yin Z, Sheng L, Qiao Y, Shao J (2024). "Assessment of multimodal large language models in alignment with human values". arXiv preprint arXiv:2403.17830.
35. a, b, cHu H, Zhang J, Zhao M, Sun Z (2023). "Ciem: Contrastive instruction evaluation method for better instruction tuning". arXiv preprint arXiv:2309.02301.
36. ^Maharana A, Kamath A, Clark C, Bansal M, Kembhavi A (2023). "Exposing and addressing cross-task inconsistency in unified vision-language models". arXiv preprint arXiv:2303.16133.
37. ^Liang VW, Zhang Y, Kwon Y, Yeung S, Zou JY (2022). "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning". Advances in Neural Information Processing Systems. 35: 17612–17625.
38. ^Byun J, Kim D, Moon T (2024). "MAFA: Managing False Negatives for Vision-Language Pre-training". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 27314–27324.

39. ^Ma R, Jin L, Liu Q, Chen L, Yu K (2020). "Addressing the polysemy problem in language modeling with attentional multi-sense embeddings." In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. pp. 8129--8133.

40. ^Ding W, Van Noord N. "IMP: Benchmarking Image Polysemy in Vision-Language Models".

41. ^Bordes F, Pang RY, Ajay A, Li AC, Bardes A, Petryk S, Mañas O, Lin Z, Mahmoud A, Jayaraman B, et al. An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247. 2024.

42. ^Li Z, Liu D, Zhang C, Wang H, Xue T, Cai W (2024). "Enhancing Advanced Visual Reasoning Ability of Large Language Models". arXiv preprint arXiv:2409.13980.

43. ^Chen L, Zhao H, Liu T, Bai S, Lin J, Zhou C, Chang B (2025). "An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models." In: European Conference on Computer Vision. Springer. pp. 19–35.

44. a, b, cMin K, Kim M, Lee K, Lee D, Jung K (2024). "Mitigating Hallucinations in Large Vision-Language Models via Summary-Guided Decoding". arXiv preprint arXiv:2410.13321.

45. a, b, cWoo S, Kim D, Jang J, Choi Y, Kim C (2024). "Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models". arXiv preprint arXiv:2405.17820.

46. ^Zhou DW, Zhang Y, Ning J, Ye HJ, Zhan DC, Liu Z (2023). "Learning without forgetting for vision-language models". arXiv preprint arXiv:2305.19270.

47. ^Huang W, Liang J, Shi Z, Zhu D, Wan G, Li H, Du B, Tao D, Ye M (2024). "Learn from Downstream and Be Yourself in Multimodal Large Language Model Fine-Tuning". arXiv preprint arXiv:2411.10928.

48. ^Zhu T, Liu Q, Wang F, Tu Z, Chen M (2024). "Unraveling Cross-Modality Knowledge Conflicts in Large Vision-Language Models". arXiv preprint arXiv:2410.03659.

49. a, b, cGhosh S, Evuru CKR, Kumar S, Tyagi U, Nieto O, Jin Z, Manocha D (2024). "Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs". arXiv preprint arXiv:2405.15683.

50. a, b, cJiang C, Xu H, Dong M, Chen J, Ye W, Yan M, Ye Q, Zhang J, Huang F, Zhang S (2024). "Hallucination augmented contrastive learning for multimodal large language model". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 27036–27046.

51. ^Cha S, Lee J, Lee Y, Yang C (2024). "Visually Dehallucinative Instruction Generation." In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. pp. 5510–5514.

52. ^Zhang J, Wang T, Zhang H, Lu P, Zheng F (2025). "Reflective Instruction Tuning: Mitigating Hallucinations in Large Vision-Language Models". In: European Conference on Computer Vision. Springer. pp. 196

–213.

53. a, b Tang J, Lin C, Zhao Z, Wei S, Wu B, Liu Q, Feng H, Li Y, Wang S, Liao L, et al. (2024). "TextSquare: Scaling up Text-Centric Visual Instruction Tuning". arXiv preprint arXiv:2404.12803.

54. ^ Park D, Qian Z, Han G, Lim SN (2024). "Mitigating dialogue hallucination for large multi-modal models via adversarial instruction tuning". arXiv preprint arXiv:2403.10492.

55. ^ Liu Y, Cao Y, Gao Z, Wang W, Chen Z, Wang W, Tian H, Lu L, Zhu X, Lu T, et al. (2024). "Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity". Science China Information Sciences. 67 (12): 1–16.

56. a, b Yu T, Yao Y, Zhang H, He T, Han Y, Cui G, Hu J, Liu Z, Zheng H, Sun M, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 13807-13816.

57. a, b Zhao Z, Wang B, Ouyang L, Dong X, Wang J, He C (2023). "Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization". arXiv preprint arXiv:2311.16839.

58. a, b Gunjal A, Yin J, Bas E (2024). "Detecting and preventing hallucinations in large vision language models". Proceedings of the AAAI Conference on Artificial Intelligence. 38 (16): 18135–18143.

59. ^ Xiao W, Huang Z, Gan L, He W, Li H, Yu Z, Jiang H, Wu F, Zhu L (2024). "Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback". arXiv preprint arXiv:2404.14233.

60. ^ Yu Q, Li J, Wei L, Pang L, Ye W, Qin B, Tang S, Tian Q, Zhuang Y (2024). "Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12944–12953.

61. ^ Chen Z, Zhu Y, Zhan Y, Li Z, Zhao C, Wang J, Tang M (2023). "Mitigating hallucination in visual language models with visual supervision". arXiv preprint arXiv:2311.16479.

62. ^ Wang L, He J, Li S, Liu N, Lim E (2024). "Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites". In: International Conference on Multimedia Modeling. Springer. p. 32–45.

63. ^ Ben-Kish A, Yanuka M, Alper M, Giryes R, Averbuch-Elor H. Mitigating open-vocabulary caption hallucinations. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. p. 22680-22698.

64. ^ Li L, Xie Z, Li M, Chen S, Wang P, Chen L, Yang Y, Wang B, Kong L (2023). "Silkie: Preference distillation for large visual language models". arXiv preprint arXiv:2312.10665.

65. ^Xie Y, Li G, Xu X, Kan M (2024). "V-DPO: Mitigating Hallucination in Large Vision Language Models vi a Vision-Guided Direct Preference Optimization". arXiv preprint arXiv:2411.02712.

66. ^Wang C, Gan Y, Huo Y, Mu Y, Yang M, He Q, Xiao T, Zhang C, Liu T, Du Q, et al. RoVRM: A Robust Visual Reward Model Optimized via Auxiliary Textual Preference Data. arXiv preprint arXiv:2408.12109. 2024.

67. a, bChen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L, et al. Internvl: Scaling u p vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 24185-24198.

68. a, bYou H, Zhang H, Gan Z, Du X, Zhang B, Wang Z, Cao L, Chang SF, Yang Y (2023). "Ferret: Refer and g round anything anywhere at any granularity". arXiv preprint arXiv:2310.07704.

69. a, bJain J, Yang J, Shi H (2024). "Vcoder: Versatile vision encoders for multimodal large language model s". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 27992–28002.

70. a, bZhang L, Yang B, Liu Q, Ma Z, Zhang S, Yang J, Sun Y, Liu Y, Bai X (2024). "Monkey: Image resolutio n and text label are important things for large multi-modal models". Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition. pages 26763–26773.

71. ^Dai W, Liu Z, Ji Z, Su D, Fung P (2022). "Plausible may not be faithful: Probing object hallucination in v ision-language pre-training". arXiv preprint arXiv:2210.07688.

72. a, bChen J, Pan Y, Li Y, Yao T, Chao H, Mei T (2023). "Retrieval augmented convolutional encoder-deco der networks for video captioning". ACM Transactions on Multimedia Computing, Communications and Applications. **19** (1s): 1–24.

73. a, bRamos R, Martins B, Elliott D (2023). "LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting". arXiv preprint arXiv:2305.19821.

74. a, bRamos R, Martins B, Elliott D, Kementchedjhieva Y (2023). "Smallcap: lightweight image captioning prompted with retrieval augmentation". Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition. pp. 2840–2849.

75. a, bSarto S, Cornia M, Baraldi L, Nicolosi A, Cucchiara R (2024). "Towards retrieval-augmented architec tures for image captioning". ACM Transactions on Multimedia Computing, Communications and Applic ations. ACM New York, NY.

76. a, bYang D, Cao B, Chen G, Jiang C (2024). "Pensieve: Retrospect-then-compare mitigates visual halluci nation". arXiv preprint arXiv:2403.14401.

77. a, b Zhao L, Deng Y, Zhang W, Gu Q (2024). "Mitigating object hallucination in large vision-language models via classifier-free guidance". arXiv preprint arXiv:2402.08680.

78. a, b Li W, Huang Z, Li H, Lu L, Lu Y, Tian X, Shen X, Ye J (2024). "Visual Evidence Prompting Mitigates Hallucinations in Multimodal Large Language Models".

79. a, b Zhao Y, Li Z, Jin Z, Zhang F, Zhao H, Dou C, Tao Z, Xu X, Liu D (2023). "Enhancing the Spatial Awareness Capability of Multi-Modal Large Language Model". arXiv preprint arXiv:2310.20357. Available from: https://arxiv.org/abs/2310.20357.

80. a, b Qu X, Chen Q, Wei W, Sun J, Dong J (2024). "Alleviating hallucination in large vision-language models with active retrieval augmentation". arXiv preprint arXiv:2408.00555.

81. a, b Woo S, Jang J, Kim D, Choi Y, Kim C (2024). "RITUAL: Random Image Transformations as a Universal Anti-hallucination Lever in LVLMs". arXiv preprint arXiv:2405.17821.

82. a, b Kim J, Yeonju K, Ro YM. What if...?: Thinking Counterfactual Keywords Helps to Mitigate Hallucination in Large Multi-modal Models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024. p. 10672–10689.

83. a, b Liu S, Ye H, Zou J (2024). "Reducing hallucinations in vision-language models via latent space steering". arXiv preprint arXiv:2410.15778. Available from: https://arxiv.org/abs/2410.15778.

84. a, b Chen J, Zhang T, Huang S, Niu Y, Zhang L, Wen L, Hu X (2024). "ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models". arXiv preprint arXiv:2411.15268.

85. ^ Fieback L, Spiegelberg J, Gottschalk H (2024). "MetaToken: Detecting Hallucination in Image Descriptions by Meta Classification". arXiv preprint arXiv:2405.19186. Available from: https://arxiv.org/abs/2405.19186.

86. a, b Gong X, Ming T, Wang X, Wei Z (2024). "DAMRO: Dive into the Attention Mechanism of LVLM to Reduce Object Hallucination". arXiv preprint arXiv:2410.04514. Available from: https://arxiv.org/abs/2410.04514.

87. a, b Leng S, Zhang H, Chen G, Li X, Lu S, Miao C, Bing L (2024). "Mitigating object hallucinations in large vision-language models through visual contrastive decoding." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882.

88. a, b Wang X, Pan J, Ding L, Biemann C (2024). "Mitigating hallucinations in large vision-language models with instruction contrastive decoding". arXiv preprint arXiv:2403.18715.

89. a, b *Kim J, Kim H, Kim Y, Ro YM (2024). "CODE: Contrasting Self-generated Description to Combat Hallucination in Large Multi-modal Models". arXiv preprint arXiv:2406.01920. Available from: https://arxiv.org/abs/2406.01920.*

90. a, b *Liu S, Zheng K, Chen W. "Paying more attention to image: A training-free method for alleviating hallucination in lvlms, 2024". arXiv. Available from: https://arxiv.org/abs/2407.21771.*

91. a, b *Huang Q, Dong X, Zhang P, Wang B, He C, Wang J, Lin D, Zhang W, Yu N (2024). "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 13418–13427.*

92. ^ *Han Z, Bai Z, Mei H, Xu Q, Zhang C, Shou MZ (2024). "Skip\n: A simple method to reduce hallucination in large vision-language models". arXiv preprint arXiv:2402.01345.*

93. ^ *Yue Z, Zhang L, Jin Q (2024). "Less is more: Mitigating multimodal hallucination from an eos decision perspective". arXiv preprint arXiv:2402.14545.*

94. ^ *Zhu L, Ji D, Chen T, Xu P, Ye J, Liu J (2024). "Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding". arXiv preprint arXiv:2402.18476.*

95. ^ *Chen Z, Zhao Z, Luo H, Yao H, Li B, Zhou J (2024). "Halc: Object hallucination reduction via adaptive focal-contrast decoding". arXiv preprint arXiv:2403.00425.*

96. ^ *Feng M, Tang Y, Zhang Z, Xu C (2024). "Do More Details Always Introduce More Hallucinations in LVLM-based Image Captioning?" arXiv preprint arXiv:2406.12663.*

97. ^ *An W, Tian F, Leng S, Nie J, Lin H, Wang Q, Dai G, Chen P, Lu S (2024). "AGLA: Mitigating Object Hallucinations in Large Vision-Language Models with Assembly of Global and Local Attention". arXiv preprint arXiv:2406.12718.*

98. ^ *Zhong W, Feng X, Zhao L, Li Q, Huang L, Gu Y, Ma W, Xu Y, Qin B (2024). "Investigating and mitigating the multimodal hallucination snowballing in large vision-language models". arXiv preprint arXiv:2407.00569. Available from: https://arxiv.org/abs/2407.00569.*

99. ^ *Favero A, Zancato L, Trager M, Choudhary S, Perera P, Achille A, Swaminathan A, Soatto S. Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 14303–14312.*

100. ^ *Wang C, Chen X, Zhang N, Tian B, Xu H, Deng S, Chen H (2024). "MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation". arXiv preprint arXiv:2410.11779.*

101. ^Deng A, Chen Z, Hooi B (2024). "Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding". arXiv preprint arXiv:2402.15300.

102. a, bZhou Y, Cui C, Yoon J, Zhang L, Deng Z, Finn C, Bansal M, Yao H (2023). "Analyzing and mitigating o bject hallucination in large vision-language models". arXiv preprint arXiv:2310.00754.

103. a, bYin S, Fu C, Zhao S, Xu T, Wang H, Sui D, Shen Y, Li K, Sun X, Chen E (2023). "Woodpecker: Hallucinat ion correction for multimodal large language models". arXiv preprint arXiv:2310.16045.

104. ^Yu CE, Jalaian B, Bastian ND (2024). "Mitigating Large Vision-Language Model Hallucination at Post -hoc via Multi-agent System". Proceedings of the AAAI Symposium Series. **4** (1): 110-113. doi:10.1609/a aaiss.v4i1.31780.

105. ^Wu J, Liu Q, Wang D, Zhang J, Wu S, Wang L, Tan T (2024). "Logical closed loop: Uncovering object hall ucinations in large vision-language models". arXiv preprint arXiv:2402.11622.

106. ^Lee S, Park SH, Jo Y, Seo M (2023). "Volcano: mitigating multimodal hallucination through self-feedb ack guided revision". arXiv preprint arXiv:2311.07362.

107. ^Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ (2024). "Hallucination of multimodal large lan guage models: A survey". arXiv preprint arXiv:2404.18930.

108. ^Zhao H, Yang F, Lakkaraju H, Du M (2024). "Towards Uncovering How Large Language Model Works: An Explainability Perspective". arXiv e-prints. pages arXiv--2402.

109. ^Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M (2024). "Explainability for large la nguage models: A survey". ACM Transactions on Intelligent Systems and Technology. **15** (2): 1–38.

110. ^Zhao H, Zhao H, Shen B, Payani A, Yang F, Du M (2024). "Beyond Single Concept Vector: Modeling Con cept Subspace in LLMs with Gaussian Distribution". arXiv preprint arXiv:2410.00153. Available from: https://arxiv.org/abs/2410.00153.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.