International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Comprehensive Survey on Vision Language Models for Image Text Retrieval

Dr. Goldi Soni
Assistant professor, AUC
gsoni@rpr.amity.edu

Ch. Kirti Yadav
B.Tech CSE, AUC
chatti.yadav@s.amity.edu

Bikash Bardhan
B.Tech CSE, AUC
bikash.bardhan@s.amity.edu

Abstract

This review paper provides a comprehensive analysis of recent research developments in Vision–Language Models (VLMs) and Large Vision–Language Models (LVLMs) that enable machines to understand and reason over both visual and textual information. This study surveys 30 scientific articles written between 2023 and 2026. These articles discuss recent progress related to multimodal learning, cross-modal retrieval, visual document understanding, and multimodal reasoning systems. Some of the most significant architectures and frameworks examined in this work include CLIP-inspired models, BLIP-2, SigLIP-2, and different kinds of adapters and prompt learning techniques which increase model efficiency. Besides, the study analyzes recent developments in the field of new benchmarks and datasets that can help evaluate multimodal models on various tasks like image retrieval, visual document understanding, captioning, and visual reasoning. At the same time, important issues like hallucinations in generated outputs, weak visual grounding, culture biases, security problems, and the computational complexity of multimodal models are considered. In their studies, some researchers propose possible ways to address these problems through better multimodal alignment, token compression, contrastive learning approaches, and instruction-based feature fusion. Finally, several works describe practical applications of vision-language models in other domains such as biomedical imaging, education, autonomous systems testing, and multimodal data augmentation. Overall, this paper serves as a consolidated reference for understanding recent progress in vision–language models and guiding future research in multimodal AI systems.

Keywords : Vision–Language Models (VLMs), Multimodal Learning, Cross-Modal Retrieval, Large Language Models (LLMs), Multimodal Artificial Intelligence

I. Introduction

Vision-language models (VLMs) and large vision-language models (LVLMs) represent a major breakthrough in the field of artificial intelligence as they can help machines comprehend and reason about visual as well as text content at once. These types of models use computer vision and natural language processing algorithms to execute various activities such as image captioning, visual question answering, visual document retrieval, composed image retrieval, and multimodal reasoning. A wide range of architectures has been created in recent years that show great performance on different tasks, including zero-shot learning, multilingual reasoning, and multimodal representation alignment, for example, BLIP-2, CLIP-based models, SigLIP-2, CoLLM, and SERVAL. New datasets like ViDoRe, MERLIM, and NegBench have been created to assess the performance of these models in various scenarios, including visual reasoning and retrieval and semantic comprehension. At the same time, there are a number of challenges associated with these types of models, namely their tendency to produce hallucinations, misunderstanding negations, security concerns, cultural bias, and high computational cost.

II. Literature Review

Visual Document Retrieval (VDR) aims to retrieve relevant visual documents such as images, scanned pages, and videos based on text queries, but traditional approaches often rely on expensive, task-specific multimodal training. The paper examines the possibility of zero-shot learning on vision-document retrieval (VDR) with pre-trained Vision-Language Models (VLMs). A method named SERVAL is proposed for zero-shot vision-document retrieval based on generate-and-encode framework. Firstly, the pre-trained VLM is used to generate descriptive text corresponding to the visual document in question. Secondly, pre-trained text encoding model is used to embed the user query and generated description of visual documents into a semantic space enabling comparisons for retrieval. Notably, no extra data is required for fine-tuning or retraining the pre-trained models, thus no supervision at all. The experiments were conducted on two benchmarks datasets - ViDoRe-v2 and MIRACL-VISION. Different numbers of VLM models were used with different sizes of parameter range from 2B to 72B as well as various text encoders. Zero-shot SERVAL reached 63.4% nDCG@5 on ViDoRe-v2 outperforming many supervised VDR methods. Further experiments demonstrated that performance is highly dependent on the quality of text encoding while being less sensitive to the size of the VLM model used. [1]

This work aims to reduce reliance on manually annotated triplets in Composed Image Retrieval (CIR) while improving multimodal fusion between reference images and modification texts using Large Language Models (LLMs). The CoLLM framework proposes the synthetic triplet generation approach on-the-fly for image-caption datasets. In this way, the proposed method does not require expensive human annotations and is scalable. The new method offers a novel LLM-based joint embedding model to capture semantic relations between the reference image and the instructions on how to modify it. The framework also introduces a massive Multi-Text Composed Image Retrieval (MTCIR) dataset with many millions of pairs of images and descriptions of their transformations. Besides, two existing benchmarks, CIRR and Fashion-IQ, are improved by removing any ambiguity in the annotations. CoLLM is evaluated on several benchmarks of CIR in the supervised and zero-shot setups. The experiments reveal that the proposed framework outperforms previous works in retrieval tasks on the composed images. Thus, CoLLM exhibits strong capabilities in retrieving images based on complex visual and textual modifications. The authors also conclude that the use of synthetic data generation improves the robustness and generalization abilities of the model. Finally, the multimodal fusion based on LLMs captures semantic relations between images and text more accurately than previous methods. [2]

This survey systematically reviews advances in Vision–Language Models (VLMs) from 2018 to 2025, aiming to unify research across architectures, training strategies, prompt engineering, adapters, and evaluation benchmarks while identifying key challenges and future research directions in multimodal learning. This paper surveys the existing major pretrained VLM architectures and distinguishes two classes of such architecture - contrastive models, where the goal is alignment of visual and textual representation spaces, and generative models, where a multimodal output is generated by the model. Also, different fine-tuning techniques such as fine-tuning the entire model or parameter-efficient methods like LoRA and adapter fine-tuning are discussed. In addition, different ways of implementing prompt engineering techniques, used to achieve zero/few-shot capability, are explored, and the way in which carefully crafted prompts make a huge difference and give a chance to perform a task without the need for extensive finetuning of the pretrained model is explained. Besides, an overview of commonly used multimodal datasets, benchmarks and evaluation criteria for such datasets is provided. Finally, a summary of results regarding existing VLM techniques is made. This shows that despite some shortcomings, VLMs have evolved into large-scale modular instruction-following

systems able to accomplish quite challenging multimodal tasks. But, despite all achievements and advancements, the major problem faced today is that of datasets, bias and evaluation of the model's performance. [3]

This work aims to improve vision–language retrieval while avoiding the expensive process of fully fine-tuning large pretrained multimodal models. This technique presents Cross-Modal Adapter, which is an efficient way to boost the interactions between visual and textual data through adapters. Rather than fine-tuning the whole pretrained network, the technique incorporates lightweight adapter layers in the vision encoder and language encoder in a pretrained model. Through this method, the model learns the task-specific representations by freezing the pretrained model backbone and using few adapter parameters. In addition, the Cross-Modal Adapter technique incorporates modality-agnostic weight sharing such that implicit interactions between visual and textual features can be boosted. Therefore, in the process of training, only the adapter layer parameters are trained, and the rest of the parameters from the backbone are kept fixed, which allows retaining the pretrained knowledge while adapting to retrieval tasks. This framework is tested and evaluated through several benchmarks in the image-text and video-text retrieval tasks, namely, MSCOCO, Flickr30K, and MSR-VTT. According to experimental results, the performance of the Cross-Modal Adapter framework in retrieval tasks is competitive compared to full fine-tuning but uses far fewer trainable parameters. [4]

In this work, we seek to enhance the fine-grained capability of CLIP models in image-text retrieval while maintaining their excellent zero-shot classification ability. The existing CLIP model only emphasizes global alignment between images and texts; thus, its fine-grained capability is constrained due to lacking fine-grained region-wise guidance in learning the relationship between images and corresponding texts. To solve this problem, DCLIP is proposed as a teacher-student framework with fine-grained, region-wise guidance. In the proposed teacher model, YOLOv8 is used to obtain saliency in images, and CLIP encoders are used for encoding both image regions and text. Bidirectional cross-modal attention over region and text embeddings is applied for generating semantically-enhanced representations of images. Such representations are used by the student model, whose image encoder corresponds to a conventional CLIP image encoder (ViT-B or ViT-L), while the text encoder is left frozen to keep the semantic space of CLIP. The student learns using a combined loss, composed of contrastive, cosine embedding distillation, and anchor losses (for large models). Notably, the student is not given any information about bounding boxes so that inference efficiency remains unaltered. The proposed model demonstrates its high performance on text-to-image retrieval despite training with a relatively small amount of data; in particular, recall@1 is increased by over 20% on MSCOCO, while zero-shot classification accuracy on ImageNet is preserved at 94%. Furthermore, it was shown that early-stage checkpoints of the teacher lead to better generalization-specialization trade-off. Overall, DCLIP provides a scalable and resource-efficient method to enhance fine-grained retrieval without sacrificing efficiency or zero-shot performance. [5]

BLIP-2 is an open-source vision-language model capable of integrating pretrained vision encoders with large language models efficiently without the costly multimodal pretraining process. The key purpose of BLIP-2 is to establish a seamless connection between visual and textual modalities while ensuring superior performance in both vision-language understanding and generation tasks. The presented architecture is based on the modular system consisting of three modules, namely, the image encoder, fixed in the pretrained version, which extracts visual information from the images, the Querying Transformer, which selects the most informative visual features by using learnable query tokens, and the frozen large language model, in charge of processing textual information and generation of language. Specifically, the Q-Former is in charge of the

transformation of representations of features across the domains, and it selects those of visual features only that are the most informative. The training procedure includes two stages, namely, vision-language representation learning, which focuses on alignment of visual and textual features, and vision-to-language generative alignment, which makes interaction with the LLM easier and more efficient. This architecture was tested on various tasks including image-text retrieval, visual question answering, image captioning, and zero-shot reasoning. It was found that BLIP-2 demonstrates superior results in all tasks considered in relation to its training efficiency when compared to other multimodal architectures. Indeed, due to the freezing of both vision encoder and the LLM during training, the capabilities of these models remain intact while making interactions with the two easy. In particular, the Q-Former helps in encoding visual signals into a compact form that can be understood by the language models. [6]

This paper focuses on improving Composed Image Retrieval (CIR), a task that retrieves a target image using a reference image and a textual modification that describes how the image should change. In most cases, existing CIR approaches either employ complicated task-oriented network designs or require large quantities of annotated training data, leading to limitations in scalability. Therefore, in order to overcome these obstacles, the authors investigate if prompt learning may be used to augment pretrained vision-language models for the purposes of cross-modal information retrieval tasks. For that purpose, the paper presents a Multimodal Prompt Learning approach that leverages pretrained CLIP architectures as its foundation. Rather than altering the architecture of the model, the approach involves additional multimodal prompts in terms of image and text prompts for the visual encoder and language encoder, respectively. Thanks to this innovation, the model learns to leverage knowledge obtained from both input modalities via jointly optimizing prompt embeddings. Importantly, in this approach, the CLIP backbone remains frozen, meaning that only the learnable prompts' weights are updated during the training procedure. In other words, the framework described above is parameter-efficient and lightweight in nature. It is validated using various CIR benchmark datasets like CIRR and Fashion-IQ, both in zero-shot and supervised settings. The results demonstrate that prompt learning is more effective in CIR compared to conventional approaches. Overall, the study shows that prompt learning is a powerful and scalable alternative to complex model modifications for composed image retrieval. [7]

This paper focuses on improving Visual Document Retrieval (VDR) by effectively capturing both textual and visual information contained in document images. Typically, traditional VDR methods represent an entire document with one-vector embedding, which does not enable them to capture fine-grained structures like tables, figures, layouts, or localized text in the documents. To address this problem, the authors have developed a new multi-vector retrieval model called CoPali that is based on the late-interaction approach. In other words, CoPali uses not one but several contextualized vectors for different parts of a document image, thus capturing its fine-grained structure. In addition, query words are mapped to token-level embeddings. Then, relevance is calculated using MaxSim, a similarity function that compares embeddings of query tokens with embeddings of document regions. For learning effective multimodal representations, CoPali employs pretrained vision-language backbones along with task-specific contrastive training. The authors have also created a ViDoRe benchmark for evaluating visual document retrieval models. As shown through experiments, the proposed model significantly outperforms not only single-vector VDR models but also early-fusion VDR approaches, demonstrating the efficiency of the proposed late-interaction strategy. Furthermore, CoPali proves to be more scalable compared to existing models. Overall, CoPali establishes a new baseline for visual document retrieval and encourages further research on structured multimodal retrieval systems. [8]

ViDoRe is a benchmark that evaluates Visual Document Retrieval (VDR) models, which need to have knowledge of visual and textual contents in the documents at once. It aims to address the problem of the lack of data sets containing real-world mixed visual-textual data, e.g., layout, table, figure, and text inside those documents. First of all, the document pages are transformed into images, and a retrieval system will be evaluated through text queries, which involve understanding the document structure, its font type, and layout and visual elements. Different types of documents are included in the data set, e.g., report, presentation, and page scans; thus, it is closer to the real world compared to the previous work. Furthermore, text queries in this benchmark require the use of multimodal reasoning rather than extracted text-only approach used by the previous benchmarks. Moreover, the protocol of evaluation with metrics nDCG@k and Recall@k is also provided in ViDoRe. ViDoRe evaluates two kinds of vision-language models, i.e., single-vector and multi-vector vision-language models, including models based on CLIP. The experimental results show that there are significant shortcomings of current vision-language models on document images with rich visuals. By providing a realistic evaluation environment, ViDoRe encourages the development of more advanced VDR methods. Overall, the benchmark has become an important standard for evaluating visual document retrieval systems and supports future research in multimodal retrieval architectures. [9]

This paper proposes ITSELF, a framework designed to improve fine-grained alignment between visual and textual modalities in retrieval tasks such as Text-Based Person Search (TBPS), where recognizing subtle visual details is essential. Alignment techniques have been seen to be influenced either externally or by auxiliary modules, which may result in incorrect alignments owing to misleading cues. The ITSELF framework overcomes this limitation of existing techniques by adopting self-supervision via the use of attention maps generated by the model itself for the purpose of aligning image patches with textual tokens without the need for external influence or auxiliary modules. GRAB (Guided Representation with Attentive Bank) is the main module used in the ITSELF framework. It converts the attention maps generated in all the layers into tokens of high salience based on visual patches and textual tokens. Following this, the proposed technique learns the local alignment objective based on these tokens, but ensuring that it maintains global similarity across the two modalities. Two additional complementary modules namely MARS (Multi-layer Attention-based Representation Selection) and ATS (Adaptive Token Scheduler) have also been introduced within the ITSELF framework to achieve this objective. While the former identifies important tokens from all the transformer layers, the latter helps determine the number of tokens for each layer with decreasing emphasis from global contexts to local tokens. [10]

In general, the goal of this survey is to review existing research about applications of LLMs in ADS testing and comprehend how LLMs could be leveraged for each stage of the testing process (i.e., scenario sourcing, scenario generation, simulation execution, safety analysis). The review of the state of the art in LLM-based ADS testing is conducted through an analysis of available literature and classification of the reviewed papers depending on whether their authors leverage LLMs to perform scenario sourcing/enrichment, scenario generation, scenario optimization, test execution, or safety performance evaluation. Moreover, the role of LLMs in different stages of scenario creation (data labeling, hazard prediction, automation of scenario generation process, etc.), simulation execution, and safety performance evaluation will be highlighted. The surveyed papers will present a variety of solutions that were developed by researchers to incorporate LLMs into automated driving testing infrastructure (different tools, methodologies, etc.). As a result of the paper, it will be possible to determine advantages and disadvantages of current approaches as well as find out what challenges

remain unresolved. Overall, findings suggest that LLMs could be efficiently used to enhance the testing process of self-driving cars thanks to their ability to generate and interpret data. [11]

This paper aims to improve Composed Image Retrieval (CIR) by better understanding user intent, where a query consists of a reference image and a textual modification describing desired changes. Previous approaches have difficulties capturing visual information along with accurate understanding of the modification text, resulting in low retrieval efficiency. In this paper, the researchers introduce a new paradigm, CIR-LVLM, that exploits Large Vision–Language Models (LVLMs) for capturing better interaction of visual and textual information to generate the retrieval results. The approach relies on the Connector module where the generated sentence prompts help the model interpret image features at sentence-level representation. Another core component of the proposed framework is Hybrid Intent Instruction, where a set of task-level prompts representing the objective of the retrieval operation and a set of instance-specific soft prompts are extracted from the learnable prompt pool that depends on both the image and instruction input. The framework employs a one-pass encoder for efficient multi-modal inference. Training is conducted under the contrastive learning setup, where the embedding of queries needs to match embeddings of correct target images. The model is validated on the standard benchmarks of CIR, i.e., Fashion-IQ, Shoes, and CIRR, using the recall-based metric. Results prove the superiority of CIR-LVLM compared to previous fusion-based and textual inversion-based models. Overall, the approach shows strong potential for future multimodal retrieval systems that require precise understanding of user intent while maintaining efficient inference. [12]

This paper addresses the shortage of large-scale biomedical image–caption datasets needed for training effective vision–language models (VLMs) in medical domains. The authors present BIOMEDICA, which consists of millions of image–caption pairs collected from publically available scientific publications. The dataset spans many modalities of medical images, ranging from radiological images, to microscopic images, to pathological images, to ophthalmic images, to dermatological images. Therefore, this dataset can be used to develop domain general biomedical vision–language models (VLM). To make sure that the data used for training biomedical VLMs is high-quality, a set of preprocessing and filtering steps have been employed in order to verify that each image–caption pair has valid caption content and usable image format. The authors use the dataset to pre-train several biomedical VLMs by training the models with similar contrastive learning objective as CLIP, hence ensuring proper alignment of image representations and textual representations. The proposed models are then used on downstream biomedical tasks such as image-text retrieval, zero-shot image classification, and biomedical visual reasoning. It is observed that models trained using the biomedical image–caption pairs perform significantly better than the general-domain VLMs. This result indicates the importance of large domain-specific biomedical image–caption dataset. Furthermore, it is found that these models exhibit strong generalization ability in biomedical tasks and across biomedical imaging modalities. [13]

This paper proposes DCLIP, a dynamic contrastive learning framework designed to improve fine-grained alignment between visual regions and textual tokens in vision–language models. However, conventional contrastive learning approaches such as CLIP primarily concentrate on global image-text matching while disregarding local correlations and object attributes. To counteract the problem, DCLIP presents a Dynamic Matching Strategy that dynamically reassigns negative samples depending on their semantic similarities in training. As a result, the model will concentrate on more challenging instances to learn and enhance its capacity to differentiate between dissimilar examples. In addition, the proposed framework includes token-to-token and token-to-

region alignment, which allows the model to discover precise correlations between textual tokens and particular regions. Also, a dynamic temperature and weighting scheme is introduced to prioritize difficult samples for training while preserving stable optimization. DCLIP is extensively evaluated across various vision-language applications, including image-text retrieval, composed image retrieval, and fine-grained matching datasets. The experimental findings reveal that DCLIP dramatically outperforms traditional CLIP and other static contrastive learning approaches on retrieval-based tasks. With dynamic contrastive guidance, the model can better align vision-language pairs at a fine-grained level and develop robust multimodal representations. It is critical to note that no additional inference overhead is incurred using dynamic contrastive learning, making it more efficient for practical deployments. [14]

This paper addresses catastrophic forgetting in vision–language models (VLMs) such as CLIP during class-incremental learning, where models must learn new classes sequentially without losing knowledge of previously learned ones. In this paper, the authors develop a PROOF framework which can keep the past knowledge and learn efficiently. As mentioned above, the image and text encoders in this work remain frozen so that the learned knowledge will not be modified. In addition, the authors add task-specific projection layers that can be expanded as new tasks arise; specifically, once a new task arises, the authors will freeze the previous task-specific projections so as to avoid catastrophic forgetting. The outputs of the task-specific projections will then be aggregated by a projection aggregation strategy. Another key component of PROOF framework is its cross-modal fusion network which contains a self-attention mechanism to fuse the projected visual feature embedding, textual class embedding, visual class prototype, and task-specific context prompt. During the process of inference, multi-branch matching will be performed to match the projected feature, visual prototype, and text class respectively. The proposed framework is extensively tested on nine datasets under different settings where the models were trained using CLIP with different pretrained weights. The experimental results demonstrate that PROOF framework indeed can alleviate catastrophic forgetting and achieve satisfactory performance in new tasks. Overall, PROOF achieves consistent improvements over vision-only continual learning methods and other VLM adaptation approaches, demonstrating a parameter-efficient and scalable solution for continual multimodal learning. [15]

This paper studies how Vision–Language Models (VLMs) process visual information when generating language from images and identifies which components are most important for vision-to-language reasoning. In their work, the authors explore information flow from image tokens to query text tokens, generated tokens, and back in VLMs like InternVL2-76B, validated by LLaVA-7B. To achieve their goals, they carry out attention knockout experiments, whereby they interrupt the information flow between various kinds of tokens and explore how visual knowledge transfers in their models. Attention distribution across all layers helps them identify which layers make more contributions towards multimodal reasoning. An evaluation scheme of using LLMs as judges measures object preservation and hallucinations in generated descriptions. In addition, the authors rely on tools for object segmentation to measure spatial localization ability captured by attention. According to their findings, query text tokens store necessary visual semantics required in generating correct descriptions. Middle transformer layers (layers 20 to 40) are most crucial for visual-to-language transfer of visual knowledge. Moreover, image tokens offer detailed visual cues about object features and locations particularly in middle transformer layers. Additionally, the researchers discovered considerable redundancy in visual tokens, allowing substantial token compression with little impact on model performance. This observation forms the basis for proposing Image

Re-prompting that reuses compressed visual context for several queries. [16]

This paper provides a comprehensive survey of security vulnerabilities in Large Vision–Language Models (LVLMs). The aim here is to distinguish between different types of attacks, examine challenges arising from multimodal models, and provide information on available attack tools, datasets, and evaluation metrics. The authors perform a comprehensive review of contemporary LVLM attacks and introduce a taxonomy that divides the attacks into four categories: adversarial attacks, jailbreak attacks, prompt injection attacks, and data poisoning/backdoor attacks. The survey considers attacks in white-box, gray-box, and black-box attacks and discusses how different attack settings affect attackers' abilities to abuse LVLM vulnerabilities. It provides insights into the use of different attack tools, datasets, LVLM architectures, and evaluation metrics like Attack Success Rate (ASR). Comparisons between attacks and defenses are conducted. According to the study, LVLM models are vulnerable because of their big size, multimodality, and interaction via instructions that create various attack surfaces. Researchers proved how attacks can be launched to influence LVLM operation to obtain potentially dangerous responses or even break the safety constraints via jailbreak techniques. Nevertheless, today, most of the attacks lack applicability and fail to generalize. More black-box attacks, cross-modal adversarial approaches, and data-centric techniques accounting for LVLM biases should be developed. Overall, the paper serves as an important reference for advancing security research and defense strategies in multimodal AI systems. [17]

This paper addresses hallucination problems in Large Vision–Language Models (LVLMs), where models generate objects or facts that are not present in the image due to strong language priors. Our objective here is to mitigate text-biased hallucinations during the autoregressive decoding process and to make the model more grounded in images without costly re-training. We present Image-Biased Decoding (IBD), which is a decoding-time approach that exploits the discrepancy between the predictions generated by the original LVLM and its image-biased variant to select less text-biased tokens. The latter is obtained by tuning the attention weights to assign higher probability to visual tokens. To select tokens grounded in visual features, we compute the decoding scores by comparing the logits produced by the original LVLM and the image-biased LVLM using a contrastive loss. For this reason, we also employ statistical techniques for discriminating content words from function words to not affect the grammaticality of generated text. We use a data-efficient and adaptive balancing scheme for switching between standard decoding and image-biased decoding depending on the Jensen–Shannon divergence between the probability distributions generated by the two LVLM variants and the type of token. We also introduce lightweight prompt tuning and plausibility constraints. We evaluate IBD on various LVLMs, such as InstructBLIP, MiniGPT-4, LLaVA-1.5, and Shikra using CHAIR and GPT-based evaluation metrics.Overall, the study demonstrates that decoding-time interventions can effectively improve the reliability and factual grounding of LVLM outputs. [18]

This paper surveys recent advances in LLM-based data augmentation, focusing on how multimodal large language models generate synthetic data for image, text, and speech modalities. The purpose of this work is to fill in the gaps in previous surveys, which mostly concentrated on augmenting data with a single modality or employing machine learning techniques. To achieve this, the authors perform a systematic literature review of scholarly papers published in prominent databases like DBLP, Google Scholar, IEEE Xplore, ACM Digital Library, PubMed, Elsevier, Nature, and Scopus. Based on their analysis, data augmentation methods can be classified into three generations: traditional data augmentation (1990-2010), machine learning/deep learning (2010-2020), and multimodal LLM-based data augmentation methods (after 2020). This survey examines 24

image augmentation papers, 45 text augmentation papers, and 35 speech augmentation papers, offering process diagrams, taxonomies, and comparative studies of each modality. It was discovered that multimodal LLMs considerably improve synthetic data generation by merging multiple sources, such as images, text, and audio, which increases semantic information and data diversity. This allows the LLMs to produce higher-quality data augmentation with richer semantics, especially in situations where data availability is scarce. Nevertheless, various obstacles have been found in LLMs' use for data augmentation, including semantic mismatch, hallucinations, prompt dependence, high computation costs, overfitting dangers, and bias and fairness issues. [19]

This paper aims to improve Vision–Language Models (VLMs) for tasks involving high-resolution and text-rich images, while reducing inference latency. In their work, the authors study the efficiency tradeoffs between resolution, vision encoder latency, visual tokens, and LLM dimensions. As an outcome, the researchers propose FastVLM based on a novel hybrid convolutional-transformer vision encoder, FastViTHD. This approach combines hierarchical downsampling, multiscale feature fusion, limited usage of self-attentive layers at high resolutions, and aggressive token compression to produce up to 16× less visual tokens compared to other vision transformers. Unlike token pruning and dynamic tiling methods, FastVLM directly scales the image resolution while remaining efficient enough. FastViTHD is plugged into the LLaVA-1.5 architecture using several language models such as Vicuna and Qwen2. The proposed model was tested on GQA, TextVQA, DocVQA, POPE, SeedBench, MMVet, and MMMU datasets. The experiments showed that FastVLM achieves significant speedups in terms of vision encoding latency and time-to-first-token (TTFT). FastVLM achieves 85× faster TTFT and utilizes a vision encoder 3.4× smaller than LLaVA-OneVision operating at 1152×1152 resolution while achieving better performance. In general, these results indicate that vision encoder efficiency plays a greater role in resolving high-resolution VLM problems than scaling the LLM. The research shows that hybrid hierarchical encoders significantly outperform Vision Transformers and token-pruning methods in these situations. [20]

This paper introduces SigLIP 2, a family of multilingual vision–language encoders designed to improve semantic understanding, localization, and dense visual representation while maintaining compatibility with existing SigLIP systems. The models are trained on the WebLI dataset consisting of 10B images and 12B alt texts in 109 languages. This allows for effective vision-language learning across many languages. The training strategy uses a combination of methods, including image-text contrastive loss based on sigmoid function, captioning and localization pre-training via decoder, and self-supervised learning by self-distillation and masked prediction. Active data curation and distillation also enable improvement of small-sized models. The newly created NaFlex architectural variant lets the models work effectively with original aspect ratios of native images and support for various image resolution levels inside one model, which makes them suitable for analyzing document-like images and images with high text content. The models come in multiple sizes from ViT-B (86M parameters) up to 1B-parameter models. SigLIP 2 is thoroughly evaluated on tasks like zero-shot classification, multilingual retrieval, dense prediction, localization, and open-vocabulary object detection. It was found out that SigLIP 2 consistently performs better than the original SigLIP and other CLIP-style architectures in all cases regardless of model size and resolution level. The multi-method training strategy brings improvements in zero-shot learning performance, multilingual retrieval, and dense visual perception. [21]

This paper proposes a training-free method for compressing visual tokens in multimodal large language models (MM-LLMs) while preserving important information for downstream tasks. The

traditional token prunes only use visual features that might result in removing visual tokens that are essential for answering text questions. In this case, the authors propose a text-guided visual token recovery procedure that leverages both visual and text features during the token prunes process. First, visual tokens are pruned by calculating visual similarity between the class token and the patch token, where the main goal is to remove visual redundancy. Then, important tokens are selected through text-visual similarity between question embeddings and visual tokens. Next, a dynamic scale filtering strategy with Local Outlier Factor (LOF) is introduced to pick up meaningful tokens for different inputs. The rest of the tokens will be further clustered and merged based on the KNN strategy. It helps keep informative background tokens while avoiding redundancy. The pruned tokens are fed into LLaVA-1.5 for downstream tasks such as ScienceQA, TextVQA, VQAv2, MME, POPE, and MMBench. This method successfully shrinks the visual tokens from the original number to less than 10%, achieving comparable results. Meanwhile, the method decreases memory consumption, computational complexity, and inference latency. Compared to other pruning approaches, the proposed method achieves equal or better performance without requiring additional training or fine-tuning. Overall, the study demonstrates that combining visual and textual guidance during token compression improves efficiency and reliability in multimodal LLMs. [22]

This paper presents a systematic review of Vision–Language Models (VLMs) in formal education, examining how these models support teaching and learning across academic contexts. Through a process involving PRISMA-based literature reviews, the researchers conducted a systematic review of the studies published between 2020 and 2025. In the course of the analysis, a total of 42 papers were selected from various databases, including ACM Digital Library, Scopus, Web of Science, Engineering Village, and IEEE Xplore. The current review examines the ways in which VLMs are used at different educational levels and in various subjects. Additionally, it explores the VLM technology categories utilized in education and their roles. The analysis highlights the following roles played by the models within the learning environment: analyst (visual interpretation); assessor (students' work assessment); content curator (creating and curating educational materials); simulator (experiential learning); and tutor (feedback). The findings suggest that most educational systems use pretrained VLMs, whereas relatively few studies consider customizing them to the domain. Among the advantages, it is mentioned that there is an increase in student engagement, personalization, and accessibility of multimodal educational resources. At the same time, some challenges are identified: low-quality data; technological limitations; educators' readiness; ethics; and the absence of a framework for assessing educational effects. Overall, the study highlights the growing potential of VLMs in education while calling for stronger research methods and evaluation standards to assess their effects on learning outcomes and teaching practices. [23]

This paper investigates security vulnerabilities in Large Vision–Language Models (LVLMs) and demonstrates how multimodal jailbreak attacks can bypass their safety mechanisms. The authors find that existing approaches aimed at manipulating visual or textual inputs alone are not effective against highly aligned models. As such, they come up with a new attack strategy called Bi-modal Adversarial Prompt (BAP) in which manipulation of both visual and textual inputs occurs. Specifically, in terms of image manipulation, they adopt query agnostic image perturbation by embedding adversarial noise into the image through PGD. In terms of textual optimization, a chain-of-thought reasoning is employed by a language model to optimize the prompts in a way that would lead the LLMs to provide positive answers to prompts that failed in the previous trials. This attack strategy was tested for effectiveness on several LVLMs (i.e., LLAVA, MiniGPT-4, and InstructBLIP) as well as other commercial LVLMs including Gemini,

ChatGLM, Qwen, and ERNIE Bot in white- and black-box setups, and datasets including SafetyBench and AdvBench. ASR was used to determine the performance of the proposed attack. The results indicate that the proposed BAP surpasses previous jailbreaks in their performances, producing about 29% higher ASRs than those reported before.Overall, BAP serves as a powerful red-teaming tool for evaluating bias, safety, and adversarial robustness in LVLMs, highlighting the need for stronger multimodal security defenses. [24]

This paper evaluates how well Vision–Language Models (VLMs) understand cultural context when generating image captions. The study aims to measure whether VLMs can correctly recognize and describe culture-specific elements in images and produce culturally sensitive captions. For this purpose, the authors develop a novel metric referred to as Cultural Awareness Score (CAS) designed to verify whether the generated captions properly describe the picture, contain appropriate cultural facts and adhere to responsible AI principles. In addition, the researchers create an original benchmark dataset, referred to as MOSAIC-1.5k, comprising 1,500 culturally-rich pictures depicting dance moves, cultural symbols, mythical stories and other culture-relevant concepts. Each image comes with human-written captions verified for biases and offensive language. The performance of four VLMs was analyzed, including GPT-4 with Vision, Gemini Pro Vision, LLaVA, and OpenFlamingo, using one prompting scheme. In addition to evaluating CAS, the researchers compared the proposed metric against classical captioning evaluation metrics such as ROUGE-L as well as hallucination rate per category. The analysis indicated that none of the tested models exhibited sufficient levels of cultural awareness, achieving CAS scores less than 40%. The highest score was produced by Gemini Pro Vision (36%), followed by GPT-4 with Vision (28%), whereas LLaVA and OpenFlamingo performed considerably lower. The models achieved slightly better performance when working with real-world cultural images, including dances, but fared poorly with abstract and vector-based cultural symbols. The findings reveal that traditional captioning metrics fail to capture cultural understanding. Overall, the study highlights a significant gap in culturally aware AI capabilities and emphasizes the need for better datasets, evaluation metrics, and culturally aligned training methods. [25]

This paper introduces dino.txt, a framework designed to connect self-supervised vision models with vision–language learning by aligning the DINOv2 visual encoder with a trainable text encoder. Our approach is aimed at solving open-vocabulary vision problems like classification and segmentation without relying on costly CLIP-like end-to-end learning. This approach is based on the Locked-image Text (LiT) methodology, according to which only the text encoder and some extra layers are learned in parallel with the frozen DINOv2 vision encoder. To create a unified vision representation, we aggregate the global [CLS] token and the average-pooled patch tokens, which allows our architecture to utilize both global and local visual cues. To mitigate the domain difference between vision and text, we introduce two lightweight vision transformer layers. Training is conducted with a CLIP-inspired contrastive loss on the weakly supervised image-text pairs. In addition, our framework includes a novel data curation pipeline with the joint use of text-based balancing (WordNet and Wikipedia concepts) and visual clustering (hierarchical k-means on DINOv2 features). The global embeddings help perform zero-shot image classification and retrieval, whereas the patch-level embeddings can be used for open-vocabulary segmentation without fine-tuning. We evaluate the effectiveness of the proposed method on several benchmarks (ImageNet, COCO, Flickr30K, ADE20K, Cityscapes, PASCAL VOC, and COCO-Stuff), showing that dino.txt outperforms existing approaches in zero-shot image classification and open-vocabulary segmentation. The study demonstrates that self-supervised vision models can rival CLIP-style multimodal systems when

properly aligned with language representations. [26]

This paper evaluates Instruction-Tuned Large Vision–Language Models (IT-LVLMs) on fundamental computer vision tasks and investigates their hallucination behavior. To make systematic analysis possible, the authors propose MERLIM, which is a benchmark dataset comprising 300,664 image-question pairs based on MS-COCO, LVIS, and Visual Genome datasets. The benchmark tests the performance of different models on three tasks of image understanding: object detection, object counting, and object relationships. Semantically equivalent prompts are used to examine sensitivity to instructions and bias introduced by prompts. Predictions of the models are converted to structured representations by means of NLP tools, such as spaCy and WordNet. The authors also utilize a procedure of inpainting to remove objects from images to check if the prediction of models remains grounded in the visual content. This leads to a definition of regular hallucinations (predicting non-existent objects) and hidden hallucinations (when the predictions look like correct but actually contradict the visual content). In their study, the authors evaluate 11 instruction-tuned VLMs, namely, BLIP-2, InstructBLIP, LLaVA-1.5, MiniGPT-4, Kosmos-2, InternLM-XComposer2-VL, and Qwen-VL-Chat. While performing rather well in the language domain, the models demonstrated poor performance on simple tasks related to vision. This is because many of the models use language hacks and priors to generate predictions. Tasks such as object counting and relationship reasoning remain challenging. Overall, MERLIM exposes hidden weaknesses in IT-LVLMs and highlights the need for stronger visual grounding and improved evaluation frameworks. [27]

This paper evaluates how well Vision–Language Models (VLMs) understand negation, an important linguistic concept for applications such as image retrieval, medical diagnosis, and safety monitoring. To investigate this issue, the authors propose NegBench, a benchmark with 79,000 examples in the domain of images, videos, and medical images taken from popular image benchmarks like COCO, VOC2007, MSR-VTT, and CheXpert. This work consists of two tasks, namely, Retrieval-Neg, where VLMs must retrieve images or videos given positive as well as negative prompts, and MCQ-Neg where models have to pick between affirmative and negated captions as an answer to a multiple-choice question. The performance of several VLMs is investigated, including various CLIP-based VLMs, NegCLIP, ConCLIP, SigLIP, AIMV2, BioMedCLIP, and CONCH. Furthermore, the authors utilize PCA to visualize the space of embeddings for the purpose of identifying biases like affirmation bias, which causes the models to consider both affirmative and negated captions as similar. In order to mitigate this issue, the authors create synthetic negation datasets for fine-tuning VLMs, namely, CC12M-NegCap and CC12M-NegMCQ, through a combined contrastive and multiple choice training strategy. Most models show poor performance on negation-related tasks, highlighting their deficiencies in language reasoning skills. CLIP-like models are observed to exhibit affirmation bias with respect to captions. However, training with negation-enriched data improves performance, increasing retrieval recall and MCQ accuracy. The study concludes that data diversity and targeted supervision are crucial for improving linguistic reasoning in VLMs, and NegBench provides an important benchmark for evaluating negation understanding in multimodal AI systems. [28]

This paper improves Large Vision–Language Models (LVLMs) by utilizing visual features from multiple layers of the vision encoder instead of relying only on the final layers. The authors first conduct a layer-wise analysis using LLaVA-v1.5 across 18 benchmarks covering six task categories, showing that different layers capture different types of information. Whereas low-level features are key to accurate perception, mid-level and high-level features are essential for semantic comprehension and reasoning operations. Taking advantage of the above observation, the authors design an instruction-guided vision aggregator

(IGVA) that automatically identifies and fuses visual features under instructions. Specifically, IGVA clusters encoder layers and encodes instructions through MPNet sentence embedding. Cross-attention weight is used to assign weights to each feature cluster, yielding a weighted combination of patch and penultimate layer features. IGVA is implemented within the context of LLaVA-v1.5 with a two-step training process including pre-training and instruction tuning, assisted by an entropy regularization auxiliary loss to avoid dependence on one feature cluster. In the experiments, instruction-guided fusion significantly exceeds static fusion and instruction-agnostic fusion, as well as boosting results on MME-P, SEED-I, MMMU, and POPE tasks. Overall, the mid-level and high-level features are crucial in general, whereas low-level features are useful for perception. Overall, the study shows that instruction-aware hierarchical feature utilization significantly improves flexibility and performance of multimodal models. [29]

This paper presents a systematic survey of text data augmentation (DA) methods using Large Language Models (LLMs), focusing on improving data quality in data-scarce NLP scenarios. The paper explores research articles collected from prominent databases like arXiv, ACL Anthology, IEEE Xplore, Springer, Google Scholar, and Scopus using stringent inclusion and exclusion criteria. It offers a two-dimensional taxonomy for organizing LLM augmentation methods according to prompt complexity (no prompt, single-step prompt, multi-step prompt, and structured prompt) and retrieval complexity (no retrieval, sparse retrieval, dense retrieval, and graph/search-based retrieval). According to the taxonomy, LLM-based text augmentation can be classified into four types: simple augmentation, prompt-based augmentation, retrieval-based augmentation, and hybrid augmentation. In addition, this review studies the role of various text granularities, namely tokens, spans, sentences, passages, contexts, and documents, in text augmentation. Comparative experiments assess the comparative performance of different augmentation techniques along with other advanced techniques such as distillation, transfer learning, and quantization. LLMs enhance text augmentation by generating human-like texts using prompt guidance. Although prompt-based and hybrid augmentation methods are usually more efficient than standard augmentation algorithms, they are still dependent on prompt engineering and suffer from hallucinations. On the other hand, retrieval-based augmentation methods ensure factual grounding by integrating external information sources with text augmentation tasks. Nevertheless, several challenges exist, including hallucination management, data redundancy, computational costs, and absence of benchmarking evaluation protocols. Overall, the survey provides a comprehensive reference for researchers addressing data scarcity and improving NLP model training using LLM-driven data augmentation. [30]

III. Objectives

The major aim of this paper is to conduct a comprehensive review of Vision–Language Models (VLMs) and Large Vision–Language Models (LVLMs). In particular, the focus will be put on how VLMs operate with both visual and linguistic data, enabling multimodal perception. Besides, this paper is aimed at analyzing existing architectural solutions for the problem of creating efficient VLMs and LVLMs, such as CLIP-based approaches, BLIP-2, and other frameworks. The learning methods of VLMs and LVLMs, especially contrastive learning, prompt learning, adapter-based learning, and others will also be covered. The paper aims to investigate the capabilities of VLMs and LVLMs to solve different problems, for example, image–text retrieval, visual reasoning, and even document understanding. Moreover, this paper will be focused on the importance of benchmarking and evaluation of VLMs/LVLMs using specific datasets. The study will reveal some of the main challenges associated with the development and usage of multimodal models, including hallucinations, weak visual grounding, cultural biases, and significant resource costs. Some of

the proposed approaches will be analyzed in detail. Furthermore, the paper will provide an insight into some applications of VLMs in such areas as healthcare, education, and autonomous robotics, among others.

IV. Comparison of Past Publishers

Table 1 Comparison of 5 Past Publishers.

Sl No.	Title	Year of Publication	Proposed Objective	Methodology	Result
1	SERVAL: Zero-Shot Visual Document Retrieval with LLMs	2025	To develop a **zero-shot visual document retrieval system** using large language models without task-specific training.	Uses a **generate-and-encode pipeline** where a Vision-Language Model generates textual descriptions of document images and matches them with queries using a text encoder.	Achieved **63.4% nDCG@5 on ViDoRe-v2 benchmark**, showing strong retrieval performance without supervised training.
2	CoLLM: A Large Language Model for Composed Image Retrieval	2025	To improve **composed image retrieval** by combining textual modification queries with visual features.	Utilizes **large language models with multimodal reasoning** to interpret query instructions and align them with image representations.	Demonstrated **better retrieval accuracy** compared to traditional multimodal retrieval approaches.
3	Cross-Modal Adapter for Vision-Language Retrieval	2025	To enhance **cross-modal retrieval** between images and text by improving alignment between visual and textual representations.	Introduces a **cross-modal adapter module** integrated into pretrained vision-language models to improve feature interaction.	Improved **retrieval performance and cross-modal understanding** across several benchmark datasets.
4	BLIP-2: Bootstrapping Language-Image Pre-training with	2023	To efficiently connect **pretrained vision encoders with**	Uses a **Querying Transformer (Q-Former)** that bridges frozen	Achieves **state-of-the-art performance in vision-language**

	Frozen Image Encoders and Large Language Models		**large language models** for multimodal tasks.	image encoders and LLMs without full model retraining.	**tasks** with significantly lower training cost.
5	DCLIP: Fine-Grained Vision-Language Alignment via Dynamic Contrastive Learning	2024	To improve **fine-grained alignment between image regions and textual descriptions**.	Applies **dynamic contrastive learning** to learn stronger image-text representations.	Produces **better semantic alignment and higher accuracy** in vision-language retrieval tasks.

V. Conclusion

As evidenced by the reviewed literature, Vision-Language Models have quickly become sophisticated multimodal systems able to undertake complicated tasks involving visual recognition and language comprehension. Approaches based on contrastive learning, prompt learning, adapter training, and instruction-driven multimodal feature combination have made great strides toward enhancing model efficiency and scalability. Many models exhibit outstanding performance in several benchmarks such as image-text matching, document understanding, and multimodal inference tasks. Nevertheless, today's systems are not free from significant shortcomings, including hallucinations, poor visual grounding capabilities, absence of cultural sensitivity, vulnerability to adversarial attacks, and challenges in comprehending language phenomena like negation. Moreover, most systems necessitate considerable computational resources and huge amounts of training data. From recent studies, it is evident that efforts should be directed towards improving evaluation metrics, constructing safer architectures, and adopting efficient training approaches to guarantee real-world applications of vision-language models.

VI. Future Scope

1. Creation of stronger multimodal benchmarks for the evaluation of vision-language models' reasoning capabilities and safety concerns.

2. Work on architecture design for efficient operation with minimal computational expenses and high accuracy.

3. Research on visual grounding for decreasing hallucinations during image captioning and other reasoning operations.

4. Humanizing vision-language models by developing cultural sensitivity in models' responses and training sets.

5. Focus on increasing security to avoid attacks and jailbreak situations.

6. Discussion of hybrid learning approaches based on self-supervision and reasoning of large language models.

7. Implementation of domain-specific models to be applied in medicine, educational, or autonomous systems.

8. Studies of human-centered multimodal machine learning models.

References

[1] T. Nguyen *et al.*, "SERVAL: Zero-Shot Visual Document Retrieval with LLMs," 2025.

[2] C. Huynh, J. Yang, A. Tawari, M. Shah, S. Tran, R. Hamid, T. Chilimbi, and A. Shrivastava, "CoLLM: A Large Language Model for Composed Image Retrieval," 2025.

[3] S. Danish, A. Sadeghi-Niaraki, S. U. Khan, L. M. Dang, L. Tightiz, and H. Moon, "A Comprehensive Survey of Vision–Language Models," 2026.

[4] H. Jiang, J. Zhang, R. Huang, C. Ge, Z. Ni, S. Song, and G. Huang, "Cross-Modal Adapter for Vision–Language Retrieval," 2025.

[5] D. Csizmadia, A. Codreanu, V. Sim, V. Prabhu, M. Lu, K. Zhu, S. O'Brien, and V. Sharma, "Distill CLIP: Enhancing Image–Text Retrieval via Cross-Modal Transformer Distillation," 2025.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," 2023.

[7] Y. Zhang, W. Wang, Z. Lin, J. Zhang, and I. Essa, "Composed Image Retrieval via Multimodal Prompt Learning," 2024.

[8] T. Faysse, Q. Sabatier, H. Déjean, P. Gallinari, and T. Wolf, "CoPali: Contextualized Late Interaction for Efficient Visual Document Retrieval," 2024.

[9] T. Faysse, Q. Sabatier, H. Déjean, P. Gallinari, and T. Wolf, "ViDoRe: Visual Document Retrieval Benchmark," 2024.

[10] T. H. Nguyen, H. L. Tran, and T. D. Ngo, "ITSELF: Attention Guided Fine-Grained Alignment for Vision-Language Retrieval," 2026.

[11] Y. Zhao, J. Zhou, D. Bi, T. Mihalj, J. Hu, and A. Eichberger, "A Survey on the Application of Large Language Models in Scenario-Based Testing of Automated Driving Systems," 2026.

[12] Z. Sun, D. Jing, G. Yang, N. Fei, and Z. Lu, "Leveraging Large Vision-Language Model as User Intent-Aware Encoder for Composed Image Retrieval," 2025.

[13] A. Lozano, P. Chambon, T. Müller, P. Jean-Louis, N. Ayache, B. Scherrer, and H. Lombaert, "BIOMEDICA: An Open Biomedical Image–Caption Archive Dataset and Vision–Language Models," 2025.

[14] Y. Zhang, X. Guo, Z. Zhang, Y. Zhao, and J. Luo, "DCLIP: Fine-Grained Vision–Language Alignment via Dynamic Contrastive Learning," 2024.

[15] D. W. Zhou, Y. Zhang, Y. Wang, J. Ning, H. J. Ye, D. C. Zhan, and Z. Liu, "Learning without Forgetting for Vision-Language Models," 2025.

[16] O. Kaduri, S. Bagon, and T. Dekel, "What is in the Image? Understanding the Vision of Vision-Language Models," in *Proc. CVPR*, 2024.

[17] D. Liu, M. Yang, X. Qu, P. Zhou, Y. Cheng, and W. Hu, "A Survey of Attacks on Large Vision-Language Models," 2024.

[18] L. Zhu, D. Ji, T. Chen, P. Xu, J. Ye, and J. Liu, "IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding," in *Proc. CVPR Workshops*, 2024.

[19] R. Sapkota, S. Raza, M. Shoman, A. Paudel, and M. Karkee, "Multimodal Large Language Models for Image, Text, and Speech Data Augmentation: A Survey," 2025.

[20] P. K. Vasu *et al.*, "FastVLM: Efficient Vision Encoding for Vision-Language Models," in *Proc. CVPR*, 2024.

[21] M. Tschannen *et al.*, "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," 2025.

[22] Y. Chen, J. Xu, X. Y. Zhang, W. Z. Liu, Y. Y. Liu, and C. L. Liu, "Recoverable Compression: A Multimodal Vision Token Recovery Mechanism Guided by Text Information," in *Proc. AAAI*, 2025.

[23] J. Tian, "Vision-Language Models in Teaching and Learning: A Systematic Literature Review," 2026.

[24] Z. Ying, A. Liu, T. Zhang, Z. Yu, S. Liang, X. Liu, and D. Tao, "Jailbreak Vision-Language Models via Bi-Modal Adversarial Prompt," 2024.

[25] O. Burda-Lassen, A. Chadha, S. Goswami, and V. Jain, "How Culturally Aware Are Vision-Language Models?" 2025.

[26] C. Jose *et al.*, "DINOv2 Meets Text: A Unified Framework for Image- and Pixel-Level Vision–Language Alignment," in *Proc. CVPR*, 2024.

[27] A. Villa, J. L. Alcázar, A. Soto, and B. Ghanem, "MERLIM: Multimodal Evaluation Benchmark for Large Image–Language Models," in *Proc. CVPR Workshops*, 2024.

[28] K. Alhamoud*et al.*, "Vision-Language Models Do Not Understand Negation," in *Proc. CVPR*, 2024.

[29] X. Li *et al.*, "Instruction-Guided Fusion of Multi-Layer Visual Features in Large Vision-Language Models," 2025.

[30] Y. Chai, H. Xie, and J. S. Qin, "Text Data Augmentation for Large Language Models: A Comprehensive Survey of Methods, Challenges, and Opportunities," 2026.