Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060363

REVIEW

A Review on Vision-Language-Based Approaches: Challenges and Applications

Huu-Tuong Ho[1,#], **Luong Vuong Nguyen**[1,#], **Minh-Tien Pham**[1], **Quang-Huy Pham**[1], **Quang-Duong Tran**[1], **Duong Nguyen Minh Huy**[2] **and Tri-Hai Nguyen**[3,*]

[1]Department of Artificial Intelligence, FPT University, Danang, 550000, Vietnam
[2]Department of Business, FPT University, Danang, 550000, Vietnam
[3]Faculty of Information Technology, School of Technology, Van Lang University, Ho Chi Minh City, 70000, Vietnam

*Corresponding Author: Tri-Hai Nguyen. Email: hai.nguyentri@vlu.edu.vn

#These authors contributed equally to this work

Received: 30 October 2024; Accepted: 20 January 2025; Published: 17 February 2025

ABSTRACT: In multimodal learning, Vision-Language Models (VLMs) have become a critical research focus, enabling the integration of textual and visual data. These models have shown significant promise across various natural language processing tasks, such as visual question answering and computer vision applications, including image captioning and image-text retrieval, highlighting their adaptability for complex, multimodal datasets. In this work, we review the landscape of Bootstrapping Language-Image Pre-training (BLIP) and other VLM techniques. A comparative analysis is conducted to assess VLMs' strengths, limitations, and applicability across tasks while examining challenges such as scalability, data quality, and fine-tuning complexities. The work concludes by outlining potential future directions in VLM research, focusing on enhancing model interpretability, addressing ethical implications, and advancing multimodal integration in real-world applications.

KEYWORDS: Bootstrapping language-image pre-training (BLIP); multimodal learning; vision-language model (VLM); vision-language pre-training (VLP)

1 Introduction

The transition in Artificial Intelligence (AI) from unimodal to multimodal systems represents a revolutionary change in how machines understand and engage with data. Conventional unimodal systems perform exceptionally well when processing inputs from a single source, such as text, music, or pictures, but they frequently fail to capture the subtleties inherent in multimodal datasets [1,2]. For example, understanding a joke requires more than just processing the words; it also involves considering the tone of voice, facial expressions, and contextual factors [3]. Comprehension is enhanced by integrating multiple sensory inputs, highlighting the limitations of single-modality approaches [4].

The evolution of advanced Vision-Language Models (VLMs), such as Bootstrapping Language-Image Pre-training (BLIP) [5], is built on a rich foundation of research and innovation. Around 2018, the field witnessed a surge in interest in multimodal learning, particularly in the context of vision-language tasks. Early works explored joint embeddings of visual and textual features, leveraging techniques such as canonical correlation analysis [6] and deep neural networks [7]. The advent of large-scale image-text datasets, including Conceptual Captions [8] and Visual Genome [9], fueled further advancements, enabling the development of more powerful VLMs [10–12]. Transformer-based architectures [13] quickly gained prominence, demonstrating exceptional capabilities in capturing long-range dependencies and contextual relationships within multimodal data [14,15]. Pre-training techniques, initially successful in natural language processing [16], were adapted for vision-language tasks, leading to significant performance gains across a variety of benchmarks [10,17,18]. This evolution culminated in sophisticated models capable of understanding individual modalities and reasoning about their complex interplay, paving the way for an AI systems generation that can truly see and speak. BLIP [5], a revolutionary framework that combines verbal and visual modalities, is at the vanguard of this change. BLIP cultivates a sophisticated grasp of the interrelationships between these aspects by using pre-training approaches on text and picture data.

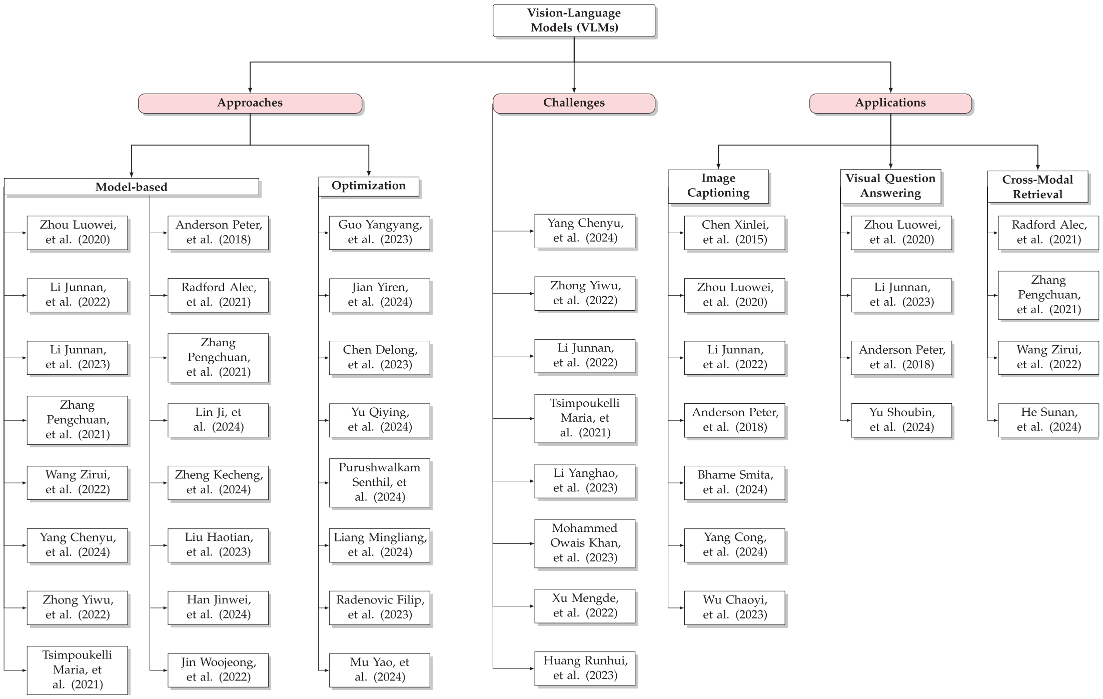

Vision-Language Models (VLMs)
Approaches
Challenges
Applications
Model-based
Optimization
Image Captioning
Visual Question Answering
Cross-Modal Retrieval
Zhou Luowei, et al. (2020)
Li Junnan, et al. (2022)
Li Junnan, et al. (2023)
Zhang Pengchuan, et al. (2021)
Wang Zirui, et al. (2022)
Yang Chenyu, et al. (2024)
Zhong Yiwu, et al. (2022)
Tsimpoukelli Maria, et al. (2021)
Anderson Peter, et al. (2018)
Radford Alec, et al. (2021)
Zhang Pengchuan, et al. (2021)
Lin Ji, et al. (2024)
Zheng Kecheng, et al. (2024)
Liu Haotian, et al. (2023)
Han Jinwei, et al. (2024)
Jin Woojeong, et al. (2022)
Guo Yangyang, et al. (2023)
Jian Yiren, et al. (2024)
Chen Delong, et al. (2023)
Yu Qiying, et al. (2024)
Purushwalkam Senthil, et al. (2024)
Liang Mingliang, et al. (2024)
Radenovic Filip, et al. (2023)
Mu Yao, et al. (2024)
Yang Chenyu, et al. (2024)
Zhong Yiwu, et al. (2022)
Li Junnan, et al. (2022)
Tsimpoukelli Maria, et al. (2021)
Li Yanghao, et al. (2023)
Mohammed Owais Khan, et al. (2023)
Xu Mengde, et al. (2022)
Huang Runhui, et al. (2023)
Chen Xinlei, et al. (2015)
Zhou Luowei, et al. (2020)
Li Junnan, et al. (2022)
Anderson Peter, et al. (2018)
Bharne Smita, et al. (2024)
Yang Cong, et al. (2024)
Wu Chaoyi, et al. (2023)
Zhou Luowei, et al. (2020)
Li Junnan, et al. (2023)
Anderson Peter, et al. (2018)
Yu Shoubin, et al. (2024)
Radford Alec, et al. (2021)
Zhang Pengchuan, et al. (2021)
Wang Zirui, et al. (2022)
He Sunan, et al. (2024)

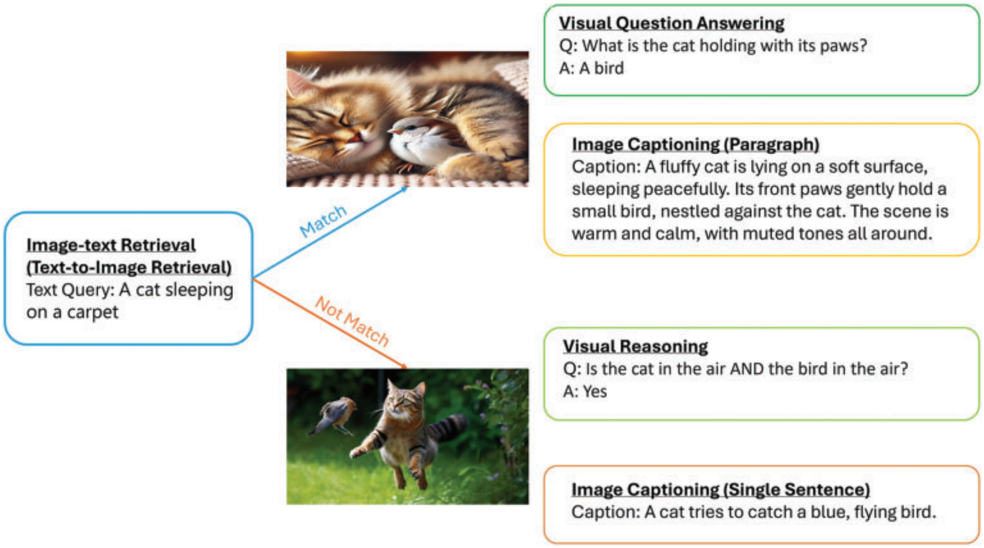
Figure 1: Taxonomy of VLMs, illustrating key approaches, challenges, and applications [5,10,19–49]

In this paper, we present an extensive review of Vision-Language Pre-training (VLP), starting with an in-depth investigation of its basic ideas and the development of important models. We look at important contributions to the area, including advancements in visual representation techniques [16] and unified frameworks for handling tasks such as image captioning and Visual Question Answering (VQA) [50]. While numerous papers have provided insights into the development of multimodal and VLP [1,4], there remains a critical gap in understanding the practical implications, which comes with an overview of the future direction of recent advancements such as BLIP [5]. This review study aims to fill the gap by offering a detailed analysis of the underlying principles and state-of-the-art approaches in VLP, highlighting the transformative impact of VLMs, and offering a focused discussion on the synthesis of vision and language and its real-world applications. To provide a better comprehensive view of VLMs, we have visualized the taxonomy as Fig. 1, which organizes the landscape of VLMs into three primary areas: approaches, challenges, and applications. Approaches such as BLIP and Contrastive Learning are designed to tackle the scalability and data quality challenges inherent in multimodal tasks. Their applicability extends to various domains, including VQA and cross-modal retrieval, highlighting their flexibility. Our contribution implies summarizing existing research and combining key insights into challenges and future potential in this domain.

The remainder of this paper is organized as follows. In Section 2, we provide a comprehensive overview of the background of VLP and explore the diverse applications of VLP models. We then analyze their performance on various benchmarks in Section 3. We discuss the challenges and future directions of VLP, addressing critical issues such as data requirements, model interpretability, and ethical considerations in Section 4. Finally, Section 5 concludes this work.

2 Overview of Vision-Language Pre-Training

VLP has garnered significant interest recently, emerging as a key methodology for bridging the gap between visual and linguistic information. At its core, VLP seeks to develop models capable of understanding and reasoning about the relationships between images and text, leading to performance improvements in diverse tasks such as image captioning, VQA, cross-modal retrieval, and even language-guided image generation, as illustrated in Fig. 2. Unifying these modalities allows for a deeper and more nuanced understanding of the world, which aligns more closely with how humans process and interpret information from multiple sources. VLP models leverage large-scale datasets containing paired image-text data to achieve this, using pre-training objectives that align textual and visual embeddings in a shared semantic space.



Visual Question Answering
Q: What is the cat holding with its paws?
A: A bird

Image Captioning (Paragraph)
Caption: A fluffy cat is lying on a soft surface, sleeping peacefully. Its front paws gently hold a small bird, nestled against the cat. The scene is warm and calm, with muted tones all around.

Image-text Retrieval (Text-to-Image Retrieval)
Text Query: A cat sleeping on a carpet

Visual Reasoning
Q: Is the cat in the air AND the bird in the air?
A: Yes

Image Captioning (Single Sentence)
Caption: A cat tries to catch a blue, flying bird.

Figure 2: An example of image-text retrieval with Visual Question Answering (VQA)

One significant milestone in VLP is the concept of unified pre-training, where models are trained to handle multiple vision-language tasks under a single framework. The work [15] introduced the Vision-and-Language BERT (ViLBERT) model, an early attempt to establish parallel streams for processing text and image data, using co-attentional transformer layers to combine visual and semantic features effectively. ViLBERT is trained with two tasks, as represented in Fig. 3, to reconstruct the input text and image, and the multimodal alignment is predicted if the description correctly describes the image. Similarly, the Visual-Linguistic BERT (VL-BERT) model [51] focuses on unifying visual and linguistic displays, emphasizing the potential for transformer-based architectures to serve as a common backbone for VLP. These approaches have laid the groundwork for subsequent advancements, where efforts have increasingly shifted toward incorporating richer visual representations, such as those found in VinVL [21]. VinVL revisits the extraction of object-level features and integrates them into VLMs, demonstrating the critical role of high-quality visual representation in enhancing model performance.

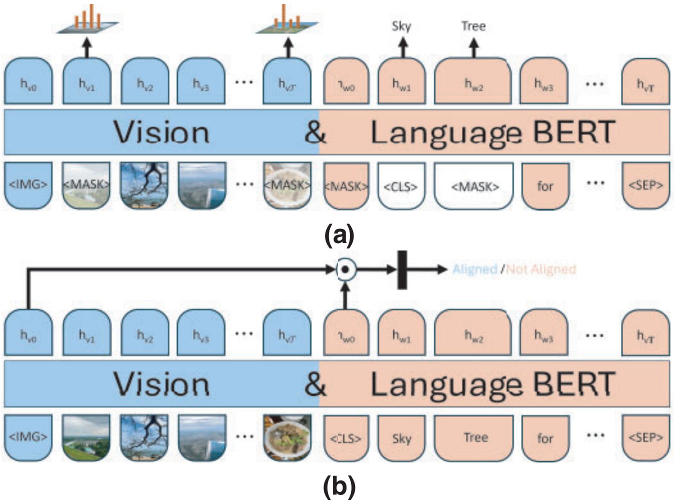


Figure 3: Two tasks in the training step of ViLBERT: (a) masked multimodal learning; (b) multimodal alignment prediction

In addition, a notable development in VLP is the use of weakly supervised learning, as seen in the Simple Visual Language Model (SimVLM) [22]. This model employs a simple, end-to-end transformer architecture, as shown in Fig. 4, and demonstrates that training on large-scale data with weak supervision, without the need for carefully curated annotations, can still achieve competitive results across various benchmarks. The success of SimVLM underscores the importance of data scalability, suggesting that larger and more diverse datasets, even when weakly labeled, can serve as a solid foundation for training robust VLP models, making this approach particularly promising for scaling VLP to broader applications.

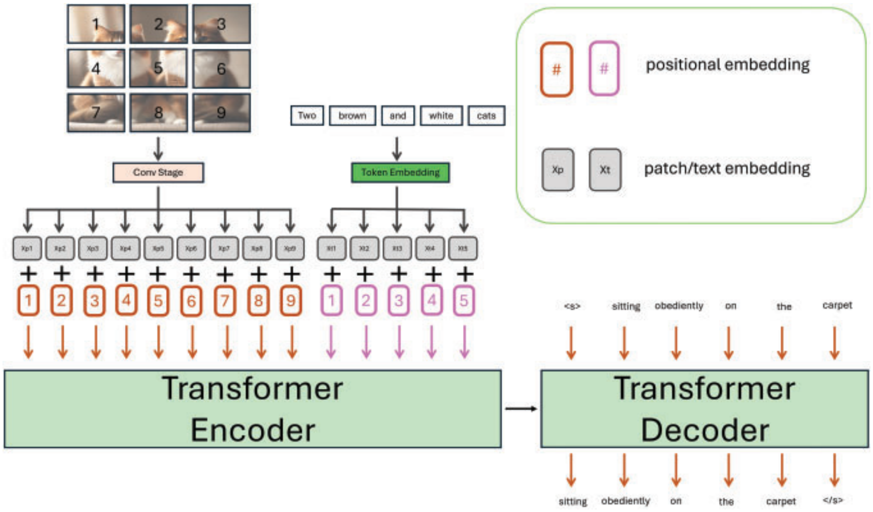


Figure 4: SimVLM Transformer architecture

Region-based Contrastive Language-Image Pre-training (RegionCLIP) [24] further enhances VLP by focusing on regional alignment between visual and textual content, which allows for a more granular level of understanding. Unlike other VLP approaches that treat the entire image as a single entity [10,11], RegionCLIP decomposes the image into meaningful regions, aligning these with corresponding linguistic descriptors, as demonstrated in Fig. 5. One prominent technique within RegionCLIP that enhances interpretability is the use of attention mechanisms, which allow the model to focus on particular parts of an image while drawing matches to relevant textual descriptors, thus providing a more detailed understanding of the content. This is particularly useful in tasks that involve complicated reasoning about object distinctions or at least regions within an image, enabling the model to produce contextually appropriate answers.

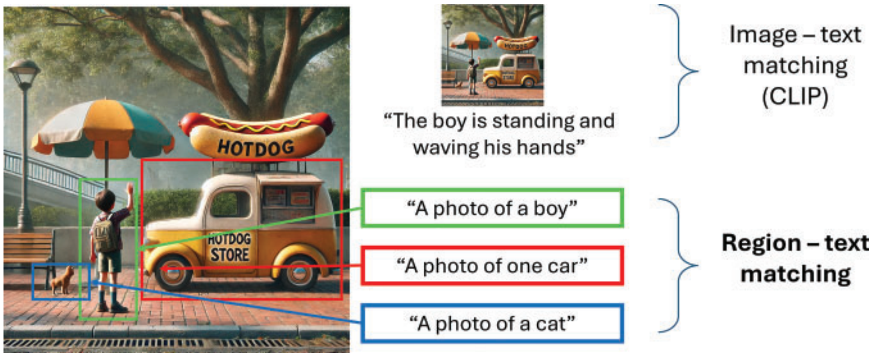


Figure 5: RegionClip and CLIP Visual-text Matching

VLP has become a nexus of innovation, weaving visual and textual modalities to enable models that recognize and describe images and infer deeper semantic alignments. Central to this effort is creating models that can seamlessly interpret image-text pairings for tasks such as image captioning and VQA, as highlighted by [19]. These systems synergize multiple vision-language tasks, emphasizing a holistic understanding of cross-modal relationships. The BLIP [5] and Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (BLIP-2) [20] frameworks push this boundary further by refining the alignment process between vision and language, introducing more robust mechanisms for harmonizing visual and textual streams under a shared semantic canopy. They embody the shift from separate pre-training streams toward a more interconnected, bidirectional understanding between modalities.

For video-based tasks, the study [52] extended VLP models to video contexts, demonstrating their ability to perform well with minimal data. This highlights the adaptability of VLP methods in transferring knowledge from static images to dynamic sequences. The work [53] further advanced this domain by showcasing the benefits of training on large video corpora. Using distillation techniques, the model efficiently extracts and retains critical visual information from vast datasets, essential for understanding complex, real-world scenarios in video format. These advancements are particularly relevant for applications such as video captioning and video content retrieval [48], where a deep understanding of video content is crucial.

The consideration of VLMs within remote sensing is either necessary or offers great potential for recent developments in multimodal systems. SpectralGPT [54] is an example of a frontier foundational model tailored to the needs of spectral remote sensing, utilizing a 3D generative pre-trained transformer (GPT) architecture. The model combines spatial-spectral information input through innovative 3D token generation and multi-target reconstruction mechanisms, enabling the processing of over a million spectral images with exceptional performance in scene classification and change detection tasks. In infrared imaging, UIU-Net [55] unveils a U-Net-based architecture particularly suited for the detection of small objects, a prerequisite for any fine-grained analysis concerning vision-language tasks. Its design retains spatial information through skip connections while improving the model's ability to detect small details that are usually missed by more traditional approaches. This closely aligns with the demands put forth by VLMs requiring accurate recognition of objects and proper context identification. LRR-Net [56] further emphasizes interpretability in hyperspectral anomaly detection. When it yields understandable information about its decision-making process, it fulfills an important requirement toward transparency in practical applications so that it entrusts and comprehends what the model output entails.

In the medical domain, MedKLIP [47] and MedDr [49] demonstrate the transformative potential of VLP in healthcare. MedKLIP integrates medical knowledge into pre-training, enhancing the interpretation of complex medical images alongside textual descriptions [47]. MedDr, on the other hand, uses diagnostic guidance to refine image-text associations, boosting performance in medical diagnostics [49]. VLMs significantly impact healthcare by improving diagnostic accuracy, assisting in medical report generation, and scaling services through automation. For instance, MedKLIP aids radiologists with context-aware image interpretations, while MedDr enhances anomaly detection in specialized tasks [47,49]. However, challenges persist, including the scarcity of annotated medical datasets, difficulties in generalizing across clinical settings, and concerns about data privacy and model explainability. Addressing these issues through privacy-preserving methods, diverse datasets, and collaborative efforts with clinicians could further solidify the role of VLMs in advancing healthcare.

The rise of large-scale internet data has driven significant progress in vision and language technologies, enabling advancements in automation. One such innovation, Manipulate-Anything [57], is an automated approach for robot manipulation in real-world environments. Unlike traditional methods, it does not rely on privileged state information, pre-defined skills, or a fixed set of objects, allowing the robot to perform a wide variety of tasks with diverse, unseen objects. This approach also facilitates behavior cloning policies that surpass the performance of human demonstrations. Similarly, recent studies on Large Language Models (LLMs) for autonomous driving have shown promise in improving planning and control systems. However, high computational demands and hallucinations remain challenges, impacting accurate trajectory prediction and control signal generation. While deterministic algorithms offer reliability, they struggle with complex driving scenarios and context-dependent uncertainty. VLM-Auto [58], a novel system for autonomous driving assistance, addresses these limitations by adapting driving behaviors based on a real-time understanding of road scenes.

The integration of VLMs in urban planning transforms the analysis and management of urban environments by combining satellite imagery and street-view visuals with textual data. UrbanVLP [59] enhances urban profiling through a dual-branch contrastive learning method that aligns visual information with generated textual descriptions, overcoming the limitations of traditional approaches. This innovative framework significantly improves the accuracy and interpretability of urban analysis, outperforming existing models in predicting key indicators such as GDP, population, and carbon emissions.

In image captioning, the paper [45] explored using VLP to generate personalized captions. This approach aligns with the broader trend of adapting VLP models for user-specific applications, demonstrating the ability to create captions that resonate with user preferences [45]. The work [46] underscored the dynamism between textual and visual inputs, especially when handling nuanced image data, e.g., remote sensing imagery. Additionally, the study [26] employed an attention-based mechanism that allows models to focus on different aspects of an image while generating captions, enhancing their descriptive capabilities.

Optimization remains a key area in VLP research. Guo et al. [32] proposed reducing the computational demands of VLP by optimizing the number of tokens required during training, enabling faster and more resource-efficient model deployment. Additionally, Jian et al. [33] introduced a method that decouples language pre-training from the vision component, allowing for focused improvements in text understanding without the need for simultaneous visual data processing. Prototypical Contrastive Language Image Pre-training (ProtoCLIP) [34] and Capsfusion [35] further optimize the scale of image-text data, improving model data efficiency. Meanwhile, BootPIG [36] pushes the boundaries of personalized image generation by refining the model's ability to generate contextually rich outputs with minimal prior data. Liang et al. [37] enhanced training efficiency by introducing tailored masking techniques. In contrast, Radenovic et al. [38] focused on the quality of training data, emphasizing the importance of high-quality negative examples for sharpening the model's ability to discern subtle variations in input data. Lastly, EmbodiedGPT [39] contributes to this dialogue by proposing methods to strengthen the model's reasoning capabilities, enhancing the interaction between visual and linguistic modalities.

The study [60] explored VLP from a multimodal translation perspective, offering a comprehensive view of the evolution of cross-modal models. These overviews highlight a collective push towards models that are more flexible and capable of distilling intricate relationships between words and visuals, adapting to increasingly complex real-world tasks [60]. In addition, the increasing use of prompt-based learning approaches [19] has introduced a new dimension to VLP, allowing models to adapt more efficiently to low-resource settings and specific downstream tasks [31]. Prompts act as targeted guidance during pre-training and fine-tuning, enhancing model flexibility and reducing the need for extensive re-training when adapting to new contexts.

Finally, datasets such as Common Objects in Context (COCO) [61], Visual Genome [9], and Flickr30k [62] have been instrumental in VLP, providing paired image-text data that serves as the foundation for pre-training many models. These benchmarks offer a consistent evaluation framework for tasks such as image captioning and VQA, enabling researchers to track progress across different models.

3 Vision-Language Models: Comparison and Discussion

The field of VLP has seen significant progress through diverse approaches and methodologies, each making unique contributions to tasks such as image captioning and VQA. These advancements can be evaluated through a comparative lens by analyzing performance on shared benchmarks while identifying each model's strengths and weaknesses.

3.1 Dataset for VLM Training

The development of VLMs is deeply reliant on large-scale, high-quality datasets that capture diverse and complex visual-textual relationships. Notable datasets include COCO Captions [44], Visual Genome [9], Flickr30k [62], and VQA 2.0 [63], each playing a distinct role in advancing VLM capabilities.

The COCO Captions [44] dataset, with its rich human-annotated descriptions, serves as a strong foundation for image captioning and image retrieval, enabling models to learn both object identification and scene-level understanding. Visual Genome [9] complements this by offering detailed region-level annotations, fostering fine-grained reasoning about object relationships, and enhancing performance in visual question answering. Similarly, Flickr30k [62] provides diverse, colloquial descriptions that aid in training models for less structured, conversational language use. The VQA 2.0 [63] dataset, through its balanced question-answer pairs, ensures unbiased training and robust reasoning capabilities for answering natural language queries based on visual inputs.

These datasets collectively enable robust representation learning, aligning visual and textual modalities for cross-modal reasoning. COCO Captions [44] and Visual Genome [9] support pre-training for object-level and scene-level comprehension, while Flickr30k [62] and VQA 2.0 [63] ensure adaptability to diverse linguistic and reasoning tasks. Additionally, Visual Genome [9] allows fine-grained object-level analysis, and VQA 2.0 [63] further refines reasoning accuracy, reducing biases and enhancing generalization to real-world scenarios. Together, these datasets underpin the versatility and effectiveness of modern VLMs.

3.2 Performance Metrics

Several key metrics are employed to assess the quality of text generated by VLMs about visual inputs. Bilingual Evaluation Understudy (BLEU) [64] measures the overlap of n-gram between the generated and reference texts. With a specific focus on sequences of up to four words, BLEU is particularly well-suited for evaluating image captioning, enabling the assessment of how closely generated captions align with human-generated references. It is defined as

$$\text{BLEU-}n = \text{BP} \cdot \exp\left(\sum_{k=1}^{n} w_k \log p_k\right), \tag{1}$$

where p_k is precision of k-grams, w_k is weights for k-gram precision, often set to $\frac{1}{n}$ for uniformity, and BP is the Brevity penalty to penalize short generated sentences, defined as

$$\text{BP} = \begin{cases} 1 & \textit{if } c > r \\ \exp\left(1 - \dfrac{r}{c}\right) & \textit{if } c \leq r \end{cases} \tag{2}$$

with c as the length of the candidate sentence and r as the length of the reference sentence.

Consensus-based Image Description Evaluation (CIDEr) [65] is also an effective evaluation metric in assessing image descriptions. It focuses on measuring the degree to which generated captions align with human-generated captions, emphasizing the consensus among human-provided captions by considering the frequency of n-grams. CIDEr is represented as

$$\text{CIDEr} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{n=1}^{4} g_n(c_i) \cdot g_n(s_i)}{\| g_n(c_i) \| \| g_n(s_i) \|}, \tag{3}$$

where N represents the number of reference captions, $g_n(c_i)$ denotes the term frequency-inverse document frequency (TF-IDF) weighted vector for n-grams of the candidate caption c_i, and $g_n(s_i)$ signifies the TF-IDF weighted vector for n-grams of the reference caption s_i.

Additionally, Metric for Evaluation of Translation with Explicit ORdering (METEOR) [66] provides a more sensitive evaluation to fluency and semantic understanding compared to BLEU. METEOR evaluates generated text by considering synonyms, stemming, and word order, with the formula as

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \gamma \cdot \text{Fragmentation Penalty}), \tag{4}$$

where F_{mean} is the Harmonic mean of precision and recall, i.e.,

$$F_{\text{mean}} = \frac{(1 + \alpha) \cdot P \cdot R}{\alpha \cdot P + R}, \tag{5}$$

with P as precision and R as recall, γ is the parameter to penalize fragmented alignments, typically set to 0.9, and the Fragmentation Penalty is Penalty for unaligned words between the candidate and reference.

Lastly, Semantic Propositional Image Caption Evaluation (SPICE) [67] focuses on the semantic content of captions, measuring how well the generated captures relationships and objects depicted in the image. SPICE is calculated by

$$\text{SPICE} = \frac{\sum_{R\epsilon S} \min\left(\text{Count}_c(R), \text{Count}_r(R)\right)}{\sum_{R\epsilon S} \text{Count}_r(R)}, \tag{6}$$

where S is the set of semantic propositions (relations, objects, and attributes) extracted from the captions, $\text{Count}_c(R)$ is the count of a proposition R in the candidate caption, and $\text{Count}_r(R)$ is the count of a proposition$R$ in the reference captions.

3.3 Comparative Analysis

Tables 1–3 provide a comparative analysis of various VLP models on three benchmark tasks: image captioning, VQA, and image retrieval. Performance is evaluated using standard datasets and evaluation metrics, such as BLEU-4, CIDEr, METEOR, and SPICE for image captioning, VQA score for visual question answering, and Recall@K (R@K) score for image retrieval. The specific details of each model are outlined below.

Table 1: Comparison of different models for Image Captioning (IS) regarding various metrics

Model	Dataset	BLEU-4	CIDEr	METEOR	SPICE
Unified VLP [19]	COCO Captions [44]	36.5	116.9	28.4	21.3
	Flickr30k [62]	30.1	67.4	23.0	17.0
VinVL [21]	COCO Captions	40.9	140.9	–	31.1
FewVLM [31]	Flickr30k	–	–	–	37.0
SimVLM [22]	COCO Captions	40.3	143.3	–	–
BLIP [5]	COCO Captions	41.7	143.5	30.0	–
RegionCLIP [24]	COCO Captions	40.5	139.2	–	–
BLIP-2 [20]	COCO Captions	43.7	123.7	–	–
	NoCaps [68]	–	–	–	–
FIBER [69]	COCO Captions	–	42.8	–	–
NLIP [43]	Flickr30K	–	–	–	135.2
LCL [23]	COCO Captions	–	87.5	–	–

Table 2: Comparison of different models for Visual Question Answering (VQA)

Model	Dataset	VQA score
Unified VLP [19]	VQA 2.0 [63]	70.3%
VinVL [21]	VQA 2.0	76.6%
FewVLM [31]	VQA 2.0 (Few-shot)	51.1%
SimVLM [22]	VQA 2.0	24.1%
BLIP [5]	VQA 2.0	77.5%
BLIP-2 [20]	VQA 2.0	79.3%
VILA [27]	VQA 2.0	80.8%

(Continued)

Table 2 (continued)

Model	Dataset	VQA score
	GQA [70]	63.3%
LCL [23]	VQA 2.0	73.4%

Table 3: Comparison of different models for Image Retrieval (IR)

Model	Dataset	R@1 score
TCL [71]	COCO [61]	62.3%
	Flickr30K [62]	88.7%
CLIP [10]	COCO	58.4%
	Flickr30K	88.0%
NLIP [43]	COCO	82.6%
Cross-Attention Transformer [72]	COCO	67.8%
	Flickr30K	88.9%
DreamLIP [28]	COCO	58.3%
	Flickr30K	87.2%

The work [26] introduced a novel combined approach to visual attention by integrating bottom-up and top-down mechanisms for tasks such as image captioning and VQA. Traditional top-down attention models focus on task-specific context to attend to predefined spatial regions, often missing object-level detail. In contrast, the study [26] proposes a bottom-up mechanism that leverages Faster R-CNN to detect objects and salient image regions, which are then weighted through top-down task-relevant context. This combined approach allows the model to focus on coarse and fine-grained details, leading to more accurate and human-like image understanding. The model sets new state-of-the-art results on the COCO Captions [44] and VQA [50] benchmarks, achieving a CIDEr score of 117.9 in image captioning and first place in the 2017 VQA Challenge with an accuracy of 70.3 on VQA 2.0 [63]. The approach enhances interpretability and performance by focusing attention at the object level, demonstrating broad applicability across vision-language tasks.

The VLP model by [19] is a prominent example of a unified approach to handling vision-language generation and understanding tasks, such as image captioning and VQA. This model is built on a shared multi-layer transformer network used for encoding and decoding, which is pre-trained on a substantial dataset of image-text pairs. The VLP model is optimized through unsupervised learning objectives, specifically bidirectional and sequence-to-sequence (seq2seq) masked vision-language prediction. The innovative aspect of this model lies in its ability to use a single architecture for two distinct types of vision-language tasks, resulting in state-of-the-art performance on benchmarks such as COCO Captions [44], Flickr30k Captions [62], and VQA 2.0 [63]. On the COCO Captions [44] test set, the VLP model achieves a BLEU-4 score of 36.5 and a CIDEr score of 116.9. These results indicate strong performance in generating captions that are both syntactically and semantically aligned with the ground truth. On the Flickr30k [62] test set, the model achieves a BLEU-4 score of 30.1 and a CIDEr score of 67.4, demonstrating its ability to generalize well across different datasets, though the performance is slightly lower than on COCO Captions [44], likely due to the differences in dataset size and complexity. However, the model's extensive pre-training requirements make it computationally intensive, posing challenges in environments with limited annotated data.

The work [10] introduced Contrastive Language-Image Pre-training (CLIP), a model that learns transferable visual representations by pre-training on 400 million image-text pairs. CLIP uses natural language supervision to align images and captions, enabling powerful zero-shot transfer to downstream tasks. On COCO [61], CLIP achieves R@1 scores of 58.4 for image-to-text retrieval and R@1 of 37.8 for text-to-image retrieval. On Flickr30K [62], it achieves R@1 scores of 88.0 for image-to-text and R@1 of 68.7 for text-to-image retrieval, demonstrating strong performance across different datasets. Additionally, CLIP matches the accuracy of ResNet-50 on ImageNet [73] in zero-shot settings without using any ImageNet [73] training data. While CLIP performs competitively across over 30 benchmarks, optimizing for task-specific scenarios is still room for improvement.

VinVL [21] significantly enhances visual feature extraction through improved object detection. By pre-training on datasets such as OpenImages, VinVL achieves robust results in tasks requiring detailed visual understanding. On the COCO Captions dataset [44], VinVL attains a BLEU-4 score of 41.0 and a CIDEr score of 140.9, outperforming many contemporaneous models. Its performance in the VQA task is also noteworthy, achieving an accuracy of 76.6 on the VQA 2.0 [63] datasets. The model's object-centric solid visual representation drives these results, although this focus may limit its effectiveness in tasks where context or abstract understanding is critical.

The research [25] introduced Frozen, a multimodal few-shot learner that extends the capabilities of large pre-trained language models to vision-language tasks by training a vision encoder to transform images into continuous embeddings. Frozen leverages this method to perform tasks such as captioning, VQA, and few-shot learning of visual categories with minimal examples. The model achieves 48.4 accuracy on VQA 2.0 [63], outpacing a blind baseline (39.1) but trailing behind specialized models such as Oscar (73.8). Frozen's advantage lies in its ability to adapt to new tasks with few examples despite not achieving state-of-the-art results. Its strength is in generalization across diverse tasks without fine-tuning, but its reliance on pre-trained language models may not perform as well on large datasets.

The study [74] proposed PICa, a method that leverages GPT-3's few-shot learning ability for knowledge-based VQA. Instead of relying on structured external knowledge bases, PICa prompts GPT-3 with image captions or tags to retrieve and reason over relevant knowledge jointly. This approach simplifies VQA by treating GPT-3 as an implicit knowledge base, achieving state-of-the-art results on the OK-VQA [75] dataset, with a significant accuracy boost to 48.0, and showing strong few-shot performance on VQA 2.0 [63].

FewVLM [31] adopts a prompt-based learning approach tailored for low-resource vision-language tasks. By leveraging prompt designs such as Prefix Language Modeling (PrefixLM) and Masked Language Modeling (MaskedLM), FewVLM efficiently guides model performance, achieving competitive results even with significantly fewer parameters. In particular, FewVLM outperforms the Frozen model (31x larger) by 18.2 in VQA tasks and achieves results comparable to PICa, which is 246x larger. This makes FewVLM especially useful for zero-shot and few-shot learning scenarios where computational resources are constrained. Despite its success, FewVLM may not match the performance of extensively pre-trained models on larger benchmarks, particularly in tasks requiring substantial prior knowledge. Nonetheless, the model's robustness in low-resource environments highlights its effectiveness and practicality.

SimVLM [22] is a minimalist approach that simplifies the pre-training process using weak supervision from large-scale noisy data. This model relies on a single Prefix Language Modeling objective and eliminates the need for object detection pre-training, simplifying the training process and improving scalability. SimVLM achieves state-of-the-art results across multiple vision-language benchmarks, including VQA [50], NLVR2, SNLI-VE, and image captioning tasks. Its strength lies in its simplicity and strong generalization capabilities, enabling zero-shot and few-shot learning. However, due to its minimalist design, its performance may be limited in tasks requiring intricate visual-textual interactions.

The paper [69] presented Fusion-In-the-Backbone-based transformER (FIBER), a vision-language model that integrates multimodal fusion directly within the backbone using cross-attention layers. This enables FIBER to handle high-level tasks such as VQA and image captioning as well as fine-grained tasks such as object detection and phrase grounding. The model adopts a two-stage pre-training strategy: i) coarse-grained pre-training on image-text data, followed by ii) fine-grained pre-training using image-text-box data. FIBER achieves state-of-the-art performance across multiple benchmarks, including 78.55 accuracy on VQA 2.0 [63], 42.8 CIDEr on COCO Captions [44], and robust results on LVIS object detection. Despite these successes, fine-grained pre-training on high-resolution images comes with significant computational costs, and the model may inherit biases from large-scale datasets.

The study [72] introduced an innovative method for Text-to-Image retrieval by embedding object priors and leveraging a cross-attention transformer. Their model improves image-text alignment and retrieval accuracy by using detected objects as anchor points. Additionally, the query-agnostic nature of the model significantly accelerates inference compared to query-dependent approaches. On the Flickr30K [62] dataset, their method achieves an R@1 score of 73.6 for Text-to-Image retrieval and 88.9 for Image-to-Text retrieval. On the COCO dataset [61], the model achieves an R@1 score of 52.4 for Text-to-Image retrieval and 67.8 for Image-to-Text retrieval, surpassing state-of-the-art methods while maintaining inference efficiency.

BLIP [5] is a novel approach that integrates vision-language understanding and generation tasks within a unified framework. This model uses a bootstrapping method, combining a multimodal mixture of encoder-decoder (MED) architecture with a captioning and filtering mechanism (CapFilt) to improve the quality of training data. BLIP achieves state-of-the-art performance across various vision-language tasks, including image-text retrieval, image captioning, VQA [50], and visual reasoning. Its strength lies in its ability to handle understanding and generation tasks effectively. Still, the complexity of its bootstrapping process and the need for large-scale data make it computationally intensive.

The following papers present various innovative approaches for advancing VLP, each focusing on distinct methodologies and evaluated on benchmarks such as ImageNet, MSCOCO, and Flickr30K. The paper [76] introduced Masked Image Pre-training on Language Assisted Representation (MILAN), which enhances masked image modeling by leveraging language-based semantic features, excelling in ImageNet classification and semantic segmentation tasks. The authors in [77] proposed MS-CLIP, which explores parameter sharing between vision and text encoders, achieving efficiency without sacrificing accuracy by sharing most transformer layers while incorporating lightweight modality-specific modules for further improvements. It achieves a 13% boost in zero-shot classification and a 1.6-point increase in linear probing across 24 downstream tasks. The study [34] proposed ProtoCLIP, which enhances CLIP through prototype-level discrimination, grouping semantically similar representations more effectively and mitigating modality gaps, improving retrieval and classification tasks. The work [78] introduced GrowCLIP, an automatic model-growing framework that scales as data increases, significantly improving zero-shot classification accuracy by 2.3% and enhancing retrieval performance. Finally, the paper [33] presented Prompt-Transformer (P-Former), decoupling language pre-training from visual components, optimizing prompt predictions with fewer image-text pairs. However, more paired data may be needed for the best results in some scenarios. Each approach has its strengths, with MS-CLIP and ProtoCLIP standing out for their performance and efficiency gains, while GrowCLIP excels in scalability and MILAN offers rich semantic feature learning.

RegionCLIP by [24] emphasizes region-specific pre-training, aligning visual regions with corresponding textual descriptions. This approach performs strongly in tasks requiring fine-grained visual understanding, such as open-vocabulary object detection and segmentation. On COCO Captions [44], RegionCLIP achieves a CIDEr score of 139.2 and a BLEU-4 score of 40.5. In open-vocabulary object detection on the LVIS [79] dataset, RegionCLIP reaches an average precision of 29.3, demonstrating its ability to generalize to unseen categories. However, relying on region-specific annotations and pre-trained models as CLIP limits its broader applicability.

The paper [71] introduced Triple Contrastive Learning (TCL) for VLP, leveraging cross-modal and within-modal self-supervision. Experimental results demonstrated that TCL delivers competitive, state-of-the-art performance on popular downstream tasks such as image-text retrieval and VQA, using datasets such as COCO [61] and Flickr30K [62]. However, the model exhibits some limitations, including biases and reduced performance for underrepresented groups.

The work [20] introduced BLIP-2, a VLP model designed to efficiently bridge the gap between visual and textual modalities by leveraging frozen pre-trained image encoders and LLMs. BLIP-2 uses a lightweight Query-Former (Q-Former), trained in two stages: vision-language representation learning and vision-to-language generative learning. The model achieves state-of-the-art performance across multiple benchmarks, such as outperforming Flamingo80B [80] by 8.7 points on zero-shot VQA 2.0 [63] while using 54 times fewer trainable parameters. BLIP-2 also achieves impressive results, including 123.7 CIDEr on NoCaps [68] and 43.7 BLEU-4 on COCO Captions [44], demonstrating strong generalization capabilities. However, BLIP-2 has limitations, including suboptimal in-context learning due to its reliance on frozen models, as evidenced by a lack of improvement in VQA tasks. Additionally, the model may produce unsatisfactory image-to-text generations due to outdated or inaccurate knowledge in the LLM, and similar to other large language models, BLIP-2 inherits the risk of generating biased or offensive content.

Huang et al. [43] introduced Noise-Robust Language-Image Pre-training (NLIP), a novel VLP framework that handles noisy image-text pairs. The framework incorporates two key strategies, noise harmonization, and noise completion, to effectively mitigate common noise issues in image-text pre-training. NLIP achieved state-of-the-art performance in tasks such as image retrieval and image captioning, with impressive results on benchmarks such as COCO [61] (R@1 of 82.6) and Flickr30K [62] (CIDEr score of 135.2).

Liu et al. [29] introduced Large Language and Vision Assistant (LLaVA), the first multimodal model to leverage GPT-4 for generating language-image instruction-following data. By instruction tuning on these data, LLaVA combines a vision encoder with an LLM to perform visual and language understanding tasks. LLaVA achieves an 85.1% relative score compared to GPT-4 on a multimodal instruction-following dataset and sets a new state-of-the-art accuracy of 92.53% when fine-tuned on ScienceQA [81], showcasing strong multimodal chat and reasoning capabilities.

The authors in [28] proposed DreamLIP, a novel VLP model that leverages long, detailed captions generated by a Multi-modality Large Language Model (MLLM) to enhance the learning of image-text alignments. By re-captioning 30 M images with longer descriptions, DreamLIP uses sub-caption sampling and a grouping loss to match sub-caption embeddings with their corresponding local image patches. This approach significantly improves the model's fine-grained representational capacity. DreamLIP outperforms state-of-the-art models, such as CLIP trained on 400 M image-text pairs, across image-text retrieval and semantic segmentation tasks, achieving an R@1 score of 87.2% on Flickr30K [62] and 58.3% on COCO [61] for text retrieval, with fewer data and better efficiency.

Lin et al. [27] introduced Visual Language (VILA), a visual language model that enhances LLMs with visual inputs by optimizing the pre-training process. The study presents several key findings. First, freezing LLMs yields decent zero-shot results, but limits in-context learning, which requires unfreezing. Second, interleaved image-text data improves performance over image-text pairs alone. Third, re-blending text-only instruction data boosts both text and visual tasks. VILA outperforms other models, achieving 80.8% on VQA 2.0 [63], 63.3% on GQA [70], 60.6% on VisWiz [82], and 66.6% on TextVQA [83] across benchmarks.

Han et al. [30] proposed Anchor-based Robust Fine-tuning (ARF) to preserve the out-of-distribution (OOD) generalization capabilities of models, particularly CLIP, during fine-tuning. ARF uses text-compensated anchors and retrieved image-text pairs to maintain rich semantic information and prevent overfitting. On the ImageNet [73] domain shift benchmark, ARF achieves an average accuracy of 61.3%, outperforming other fine-tuning methods. In zero-shot learning, ARF achieves 55.6% accuracy across tasks such as Caltech101, OxfordPets, and StanfordCars, while maintaining competitive in-distribution performance.

Finally, the study [23] proposed Latent Compression Learning (LCL), a novel framework for pre-training vision models on interleaved image-text data, leveraging the mutual information between inputs and outputs of a causal attention model. LCL achieves competitive performance compared to CLIP on the image-text pair dataset LAION-400M and significantly outperforms other methods on the interleaved dataset MMC4. It achieves 75.2% top-1 accuracy on ImageNet [73], 48.5% R@1 on COCO retrieval [61], and 87.5 CIDEr on COCO captions [44].

3.4 Summary and Discussion

In general, VLP has seen the development of diverse models, each with unique strengths and weaknesses. For example, the Bottom-Up and Top-Down Attention model [26] excels in object-level detail extraction, leading to superior performance in image captioning and VQA, but its reliance on complex object detection might limit generalizability. VLP [19] offers a unified approach for image captioning and VQA through transformer-based architectures, but its heavy pre-training requirements pose computational challenges. CLIP [10] shines in zero-shot learning across tasks, though it falls short in task-specific optimization.

VinVL [21] improves visual feature extraction, setting benchmarks in image captioning and VQA, but may struggle in abstract contextual tasks. SimVLM [22] simplifies pre-training using weak supervision, allowing scalability and robust generalization, but may underperform in tasks requiring detailed visual-text interaction. BLIP [5] integrates vision-language understanding and generation effectively, although its bootstrapping complexity demands substantial computational resources.

FewVLM [31] is particularly strong in low-resource environments, offering competitive performance with fewer parameters, though it may not excel on larger, more complex datasets. Frozen [25] demonstrates adaptability in few-shot learning but lacks state-of-the-art results compared to specialized models. DreamLIP [28] focuses on fine-grained image-text alignment, achieving top-tier performance in image retrieval tasks, although it still relies on large-scale data and processing.

4 Challenges and Future Directions

4.1 Challenges

When scaling language-image pre-training models, various challenges require substantial computational and data-handling capacities. For example, the authors in [40] faced significant difficulties when increasing the dataset from 400 million to 2 billion images using the LAION-2B dataset. This process required a massive infrastructure for data management and model training. Training VLMs on large datasets, such as LAION-2B, presents scalability challenges. Specifically, the amount of data requires precise pre-training and fine-tuning strategies to manage computational resources while avoiding overfitting issues effectively. Research in this field aims to develop approaches that decrease the complexity of training while maintaining efficacy. These include knowledge distillation and progressive resizing, which allow more effective exploitation of computational resources by concentrating on the most informative samples during training [84]. Furthermore, recent innovations in model architectures aim to make the training process smoother by introducing adaptive learning rates and dynamic sampling strategies that could dramatically scale up the process without compromising interpretability or performance. Switching from a Vision Transformer-Large (ViT-L) image encoder to a Vision Transformer-Huge (ViT-H) [11], which has twice the parameters, dramatically increased the computational load. Additionally, scaling the text encoder to match the larger image encoder exacerbated the strain on computational resources. Despite increasing the sampled data to 25.6 billion (64 epochs of 400 million data points), the longer training times did not always yield proportional performance improvements, highlighting the challenge of balancing computational cost and model effectiveness.

In the medical domain, the study [85] encountered another significant obstacle: the lack of textual descriptions for brain scans. This shortage hindered the model's ability to perform VQA effectively, as the absence of paired text made it difficult to describe observed features. Although MedBLIP aimed to mitigate these computational challenges by not training models from scratch and leveraging pre-trained VLMs, the substantial computational power required remained a challenge.

Additionally, the work [41] introduced further complications when curating large, high-quality, multilingual, multimodal datasets. Many current VLMs are predominantly trained using just one language, primarily English, which may lead to biases related to cultures and regions and restrict the use of VLMs in different linguistic contexts [10,86]. Training VLMs with multilingual text significantly improves their ability to understand diverse cultures and visual content across languages, enhancing their effectiveness in various linguistic contexts. The AltCLIP framework [87] introduces an innovative approach to training bilingual and multilingual models by integrating the XLM-R text encoder in place of OpenAI's CLIP encoder. This method employs a two-step training process that includes teacher learning and contrastive learning, which contributes to the model's overall performance. The paper [88] presents KELIP, a bilingual model that has been trained on an impressive 1.1 billion image-text pairs, comprising 708 million in Korean and 476 million in English. By utilizing advanced techniques such as MAE pre-training and multi-crop augmentation, KELIP excels in both languages. This work illustrates the potential of multilingual training and effective strategies for enhancing the capabilities of VLMs. Training for VLMs often requires large datasets and substantial computational resources, raising sustainability concerns. One possible solution is to train effective VLMs using smaller image-text datasets by leveraging the supervision among image-text pairs [89,90]. Recent studies have also investigated pre-training VLMs with LLMs [91,92], improving the process by enriching the textual data associated with image-text pairs. This additional language knowledge enhances the model's ability to learn vision-language correlations more effectively. Managing and processing vast amounts of data from multiple languages required large-scale computational resources and meticulous alignment and synchronization between image and text data across different languages and modalities. Even though there were efforts to ensure high-quality annotations, ensuring consistency across such diverse datasets continued to pose significant computational and logistical hurdles. Moreover, the complexity of combining these modalities to create a cohesive model that performs well across languages underscores the need for sophisticated systems that can handle the scale and diversity of such datasets.

The challenges related to model interpretability and explainability arise from the two-stage framework [42]. While the mask proposal generation provides visual insights into segmented regions, the challenge lies in the granularity of these proposals, as they may overlap or fail to delineate particular objects properly, making interpretation difficult in some cases. The reliance on the CLIP-based classification system also introduces limitations because the model explanations depend heavily on the textual prompts used. Crafting appropriate prompts for each class can be subjective and may not always yield the best classification, thus posing a challenge to maintaining transparency and consistency in model outputs. Moreover, while prompt learning offers a way to improve the adaptability of text prompts, tuning these prompts requires extensive computation, and even minor changes can drastically affect the model's interpretability. Balancing the accuracy of predictions with clear, interpretable outputs remains an ongoing challenge, particularly in complex scenes where multiple classes and objects interact.

4.2 Ethical Considerations

Ethical considerations in VLMs encompass crucial dimensions such as safety, privacy, equity, and the potential for harm, thereby ensuring the responsible evolution and application of these technologies. The term "ethics" in this context denotes adherence to societal norms designed to protect individuals and communities from harm, bias, and exploitation while fostering equity, security, and transparency [93,94]. Safety pertains to the creation of models that offer sound advice, circumvent the generation of harmful or misleading information, and deter misuse through strategies such as reinforcement learning from human feedback (RLHF) and supervised fine-tuning (SFT) [93,95]. Privacy concerns are centered on the protection of sensitive and identifiable information contained in multimodal datasets through stringent data governance and anonymization techniques, with frameworks such as VLSBench playing a role in preventing visual safety information leakage (VSIL) [93]. Fairness is concerned with ensuring equitable treatment across various demographics, addressing biases with diverse datasets, fairness-oriented training objectives, and alignment datasets such as SPA-VL [94]. The mitigation of harmfulness involves averting outputs that incite violence, promote inappropriate content, or disseminate harmful ideologies through adversarial training and harm detection systems [95]. Comprehensive evaluation frameworks such as VLFeedback are instrumental in examining responses for biases, offensiveness, and safety, thereby facilitating continuous enhancements in alignment and robustness [95]. Collectively, these measures provide a robust foundation for ethically sound VLMs, enabling their safe and equitable deployment in practical settings.

4.3 Research Directions

Future research directions for advancing VLMs across different tasks are highlighted in [96,97]. In particular, improvement in the generalizability and transferability of prompt learning methods should be focused on recommendations made for scaling instance-conditional prompts to larger models, using more extensive training images, and incorporating diverse datasets [96]. Further exploration of the Conditional Context Optimization approach in new tasks and domains is recommended to assess its robustness and effectiveness, aiming to develop more scalable prompt learning techniques. In [97], challenges in open-vocabulary object detection are addressed through methods that optimize embeddings for negative proposals, improving the distinction from class embeddings. Refinement of positive proposals is also suggested to enhance context grading and strengthen prompt representation learning. Extension of the DetPro approach to additional datasets is emphasized as a key step in evaluating generalization capabilities across tasks.

Future research is also encouraged to focus on advanced pseudo-labeling techniques and self-training methods to enhance unsupervised prompt learning [98]. The proposal includes extending the Unsupervised Prompt Learning (UPL) framework beyond image classification to tasks such as object detection and segmentation to validate its generalizability. Optimizing interaction between image and text encoders within VLMs is also highlighted as essential for boosting performance and efficiency. Similarly, in [43], a Noise-robust Language-Image Pre-training (NLIP) framework is proposed, introducing noise-harmonization and noise-completion schemes to stabilize pre-training. These noise-robust learning methods could be applied across various cross-modal pre-training models, potentially improving performance in fine-grained tasks such as open-world object detection, segmentation, and image generation.

Future research directions for the BLIP-2 framework suggest scaling it to integrate larger image and language models to enhance zero-shot performance, with an emphasis on improving cross-modal alignment by refining interactions between the Q-Former and language models [20]. In addition, incorporating in-context learning through sequential image-text pairs could advance performance on VQA tasks. Extending BLIP-2 to address more complex challenges, such as visual commonsense reasoning and open-world image generation, is also recommended, along with mitigating risks associated with bias, misinformation, and privacy. Lastly, in the domain of multimodal machine translation (MMT), a recent survey [60] highlights the importance of addressing weakly grounded datasets, which lack direct visual-text relevance, such as How2, in contrast to strongly grounded datasets, such as COCO. To tackle these grounding issues, the survey advocates expanding datasets through automated means and developing pre-trained models that more effectively integrate visual information into text-to-text translation, particularly for applications with weaker visual relevance, such as subtitle translation. Furthermore, the high reliance on human annotations presents an ongoing challenge in producing high-quality, large-scale datasets, emphasizing the need for scalable solutions in both BLIP-2 and MMT frameworks.

5 Conclusion

In recent years, tremendous progress has been made in the field of VLP, with models such as BLIP [5], RegionCLIP [24], and FIBER [69] exhibiting outstanding performance on a variety of tasks. These state-of-the-art methods have used cutting-edge methodologies such as coarse-to-fine pre-training, region-based learning, and unified pre-training to improve their understanding and generation skills. Consequently, VLMs have been used in various domains, including cross-modal retrieval, picture captioning, VQA, and even specialized fields such as medical report production. Even with these noteworthy successes, much work must be done before VLP models can reach their full potential. Promising prospects for future research directions are presented by persistent challenges, including the need for better model interpretability and explainability, the ever-growing data and computational requirements [99], and the crucial need to ensure ethical considerations in the development and deployment of these potent technologies [100].

Prospects for VLP model development are quite promising for novel developments. With the potential to facilitate more intuitive, natural, and contextually aware communication, these models can completely transform the way humans and machines communicate [101]. VLMs are expected to revolutionize a wide range of industries, including healthcare [47,49], education, and entertainment [19], as they become more reliable, scalable, and interpretable. The future of VLP lies in the collective efforts of researchers, developers, and stakeholders to address the remaining challenges, explore novel architectures and learning paradigms [102], and ensure the ethical and responsible development of these powerful technologies [103]. By embracing this challenge, we can unlock the true potential of VLP models and pave the way for a future where human-machine collaboration reaches unprecedented levels of sophistication and effectiveness.

Acknowledgement: This research was supported by AIT Laboratory, FPT University, Danang Campus, Vietnam 2024.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Huu-Tuong Ho, Luong Vuong Nguyen and Tri-Hai Nguyen; methodology, Huu-Tuong Ho, Luong Vuong Nguyen and Duong Nguyen Minh Huy; investigation, Huu-Tuong Ho, Minh-Tien Pham, Quang-Huy Pham and Quang-Duong Tran; writing—original draft preparation, Huu-Tuong Ho, Luong Vuong Nguyen, Minh-Tien Pham, Quang-Huy Pham and Quang-Duong Tran; writing—review and editing, Duong Nguyen Minh Huy and Tri-Hai Nguyen; visualization, Huu-Tuong Ho, Minh-Tien Pham, Quang-Huy Pham and Quang-Duong Tran; supervision, Luong Vuong Nguyen and Tri-Hai Nguyen; project administration, Luong Vuong Nguyen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Abbreviations

AI	Artificial Intelligence
BLEU	Bilingual Evaluation Understudy
BLIP	Bootstrapping Language-Image Pre-Training
BLIP-2	BLIP with Frozen Image Encoders & Large Language Models
CIDEr	Consensus-based Image Description Evaluation
CLIP	Contrastive Language-Image Pre-Training
GPT	Generative Pre-training Transformer
LLM	Large Language Model
METEOR	Metric for Evaluation of Translation with Explicit ORdering
QA	Question Answering
SPICE	Semantic Propositional Image Caption Evaluation
VLM	Vision-Language Models
VLP	Vision-Language Pre-Training
VQA	Visual Question Answering

References

1. Liang PP, Zadeh A, Morency LP. Foundations & trends in multimodal machine learning: principles, challenges, and open questions. ACM Comput Surv. 2024 Jun;56(10):1–42. doi:10.1145/3676164.
2. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proceedings of the 28th International Conference on International Conference on Machine Learning. ICML'11; 2011; Madison, WI, USA: Omnipress; p. 689–96.
3. Liebenthal E, Silbersweig DA, Stern E. The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. Front Neurosci. 2016 Nov;10:506. doi:10.3389/fnins.2016.00506.
4. Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019 Feb;41(2):423–43. doi:10.1109/TPAMI.2018.2798607.
5. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, editors. Proceedings of the 39th international conference on machine learning. Baltimore, Maryland, USA: Proceedings of Machine Learning Research; 2022; PMLR; vol. 162, p. 12888–900.
6. Andrew G, Arora R, Bilmes J, Livescu K. Deep canonical correlation analysis. In: Dasgupta S, McAllester D, editors. Proceedings of the 30th international conference on machine learning. Atlanta, GA, USA: Proceedings of Machine Learning Research; 2013; vol. 28, p. 1247–55.
7. Socher R, Chen D, Manning CD, Ng A. Reasoning with neural tensor networks for knowledge base completion. In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in neural information processing systems. Vol. 26. Lake Tahoe, Nevada, USA: Curran Associates, Inc.; 2013.
8. Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018; Melbourne, Australia: Association for Computational Linguistics.
9. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis. 2017 Feb;123(1):32–73. doi:10.1007/s11263-016-0981-7.

10. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning, ICML 2021. vol. 139. Virtual Event: Proceedings of Machine Learning Research; 2021 Jul 18–24. p. 8748–63.
11. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:201011929. 2021.
12. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20; 2020; Cambridge, MA, USA.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems. Long Beach, CA, USA: Curran Associates, Inc.; vol. 30, 2017.
14. Tan H, Bansal M. LXMERT: learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019; Hong Kong, China: Association for Computational Linguistics.
15. Lu J, Batra D, Parikh D, Lee S. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019; Red Hook, NY, USA: Curran Associates Inc.
16. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Volume 1 (Long and Short Papers); Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86.
17. Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. Visualbert: a simple and performant baseline for vision and language. arXiv:190803557. 2019.
18. Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, et al. UNITER: universal image-text representation learning. Gewerbestrasse, Cham, Switzerland: Springer International Publishing; 2020. p. 104–20.
19. Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J. Unified vision-language pre-training for image captioning and VQA. Proc AAAI Conf Artif Intell. 2020 Apr;34(7):13041–9. doi:10.1609/aaai.v34i07.7005.
20. Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th international conference on machine learning. Vol. 202. Honolulu, HI, USA: Proceedings of Machine Learning Research; 2023. p. 19730–42.
21. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, et al. VinVL: revisiting visual representations in vision-language models. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021; Los Alamitos, CA, USA: IEEE Computer Society; p. 5575–84.
22. Wang Z, Yu J, Yu AW, Dai Z, Tsvetkov Y, Cao Y. SimVLM: simple visual language model pretraining with weak supervision. In: International Conference on Learning Representations; 2022; Amherst, MA, USA.
23. Yang C, Zhu X, Zhu J, Su W, Wang J, Dong X, et al. Vision model pre-training on interleaved image-text data via latent compression learning. arXiv:240607543. 2024.
24. Zhong Y, Yang J, Zhang P, Li C, Codella N, Li LH, et al. RegionCLIP: region-based language-image pretraining. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA: IEEE.
25. Tsimpoukelli M, Menick J, Cabi S, Eslami SMA, Vinyals O, Hill F. Multimodal few-shot learning with frozen language models. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21; 2024; Red Hook, NY, USA: Curran Associates Inc.
26. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Los Alamitos, CA, USA: IEEE Computer Society; p. 6077–86.

27. Lin J, Yin H, Ping W, Molchanov P, Shoeybi M, Han S, et al. VILA: on pre-training for visual language models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; IEEE, vol. 35, p. 26679–89. doi:10.1109/CVPR52733.2024.02520.
28. Zheng K, Zhang Y, Wu W, Lu F, Ma S, Jin X, et al. DreamLIP: language-image pre-training with long captions. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 73–90.
29. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in neural information processing systems. New Orleans, LA, USA: Curran Associates, Inc; 2023. vol. 36, p. 34892–916.
30. Han J, Lin Z, Sun Z, Gao Y, Yan K, Ding S, et al. Anchor-based robust finetuning of vision-language models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; Seattle, WA, USA: IEEE; p. 26909–18.
31. Jin W, Cheng Y, Shen Y, Chen W, Ren X. A good prompt is worth millions of parameters: low-resource prompt-based learning for vision-language models. In: Muresan S, Nakov P, Villavicencio A, Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers); Dublin, Ireland: Association for Computational Linguistics; 2022. p. 2763–75.
32. Guo Y, Zhang H, Wong Y, Nie L, Kankanhalli M. Elip: efficient language-image pre-training with fewer vision tokens. arXiv:230916738. 2023.
33. Jian Y, Gao C, Vosoughi S. Bootstrapping vision-language learning with decoupled language pre-training. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23; 2024; Red Hook, NY, USA: Curran Associates Inc.
34. Chen D, Wu Z, Liu F, Yang Z, Zheng S, Tan Y, et al. ProtoCLIP: prototypical contrastive language image pretraining. IEEE Trans Neural Netw Learn Syst. 2023:1–15. doi:10.48550/arXiv.2206.10996.
35. Yu Q, Sun Q, Zhang X, Cui Y, Zhang F, Cao Y, et al. CapsFusion: rethinking image-text data at scale. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; IEEE; vol. 35, p. 14022–32. doi:10.1109/CVPR52733.2024.01330.
36. Purushwalkam S, Gokul A, Joty S, Naik N. Bootpig: bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. arXiv:240113974. 2024.
37. Liang M, Larson M. Centered masking for language-image pre-training. Gewerbestrasse, Cham, Switzerland: Springer Nature Switzerland; 2024. p. 90–106.
38. Radenovic F, Dubey A, Kadian A, Mihaylov T, Vandenhende S, Patel Y, et al. Filtering, distillation, and hard negatives for vision-language pre-training. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Vancouver, BC, Canada: IEEE; p. 6967–77.
39. Mu Y, Zhang Q, Hu M, Wang W, Ding M, Jin J, et al. EmbodiedGPT: vision-language pre-training via embodied chain of thought. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23; 2024; Red Hook, NY, USA: Curran Associates Inc.
40. Li Y, Fan H, Hu R, Feichtenhofer C, He K. Scaling language-image pre-training via masking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Vancouver, BC, Canada: IEEE.
41. Mohammed OK, Aggarwal K, Liu Q, Singhal S, Bjorck J, Som S. Bootstrapping a high quality multilingual multimodal dataset for Bletchley. In: Khan E, Gonen M, editors. Proceedings of the 14th Asian Conference on Machine Learning, Hyderabad, India: Proceedings of Machine Learning Research; 2023; vol. 189, p. 738–53.
42. Xu M, Zhang Z, Wei F, Lin Y, Cao Y, Hu H, et al. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. Gewerbestrasse, Cham, Switzerland: Springer Nature Switzerland; 2022. p. 736–53.
43. Huang R, Long Y, Han J, Xu H, Liang X, Xu C, et al. NLIP: noise-robust language-image pre-training. Proc AAAI Conf Artif Intell. 2023 Jun;37(1):926–34. doi:10.1609/aaai.v37i1.25172.
44. Chen X, Fang H, Lin T, Vedantam R, Gupta S, Dollár P, et al. Microsoft COCO captions: data collection and evaluation server. arXiv:1504.00325. 2015.

45. Bharne S, Bhaladhare P. Enhancing user profile authenticity through automatic image caption generation using a bootstrapping language–image pre-training model. In: RAiSE-2023. Basel, Switzerland: MDPI; 2024. Vol. 51, p. 182.
46. Yang C, Li Z, Zhang L. Bootstrapping interactive image-text alignment for remote sensing image captioning. IEEE Trans Geosci Remote Sens. 2024;62:1–12. doi:10.1109/TGRS.2024.3359316.
47. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. MedKLIP: medical knowledge enhanced language-image pre-training for X-ray diagnosis. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023; Paris, France: IEEE.
48. Yu S, Cho J, Yadav P, Bansal M. Self-chained image-language model for video localization and question answering. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23; 2024; Red Hook, NY, USA: Curran Associates Inc.
49. He S, Nie Y, Chen Z, Cai Z, Wang H, Yang S, et al. MedDr: diagnosis-guided bootstrapping for large-scale medical vision-language learning. arXiv:240415127. 2024.
50. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, et al. VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015; Santiago, Chile: IEEE.
51. Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, et al. Vl-bert: pre-training of generic visual-linguistic representations. arXiv:190808530. 2019.
52. Wang Z, Li M, Xu R, Zhou L, Lei J, Lin X, et al. Language models with image descriptors are strong few-shot video-language learners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22; 2024; Red Hook, NY, USA: Curran Associates Inc.
53. Zhao Y, Zhao L, Zhou X, Wu J, Chu CT, Miao H, et al. Distilling vision-language models on millions of videos. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; Seattle, WA, USA: IEEE; p. 13106–16.
54. Hong D, Zhang B, Li X, Li Y, Li C, Yao J, et al. SpectralGPT: spectral remote sensing foundation model. IEEE Trans Pattern Anal Mach Intell. 2024;46(8):5227–44. doi:10.1109/TPAMI.2024.3362475.
55. Wu X, Hong D, Chanussot J. UIU-Net: U-Net in U-Net for infrared small object detection. IEEE Trans Image Process. 2023;32:364–76. doi:10.1109/TIP.2022.3228497.
56. Li H, Xu T, Wu XJ, Lu J, Kittler J. LRRNet: a novel representation learning guided fusion network for infrared and visible images. IEEE Trans Pattern Anal Mach Intell. 2023 Sep;45(9):11040–52. doi:10.1109/TPAMI.2023.3268209.
57. Duan J, Yuan W, Pumacay W, Wang YR, Ehsani K, Fox D, et al. Manipulate-anything: automating real-world robots using vision-language models. arXiv:240618915. 2024.
58. Guo Z, Lykov A, Yagudin Z, Konenkov M, Tsetserukou D. Co-driver: VLM-based autonomous driving assistant with human-like behavior and understanding for complex road scenes. arXiv:240505885. 2024.
59. Hao X, Chen W, Yan Y, Zhong S, Wang K, Wen Q, et al. UrbanVLP: a multi-granularity vision-language pre-trained foundation model for urban indicator prediction. arXiv:240316831. 2024.
60. Gwinnup J, Duh K. A survey of vision-language pre-training from the lens of multimodal machine translation. arXiv:230607198. 2023.
61. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision–ECCV 2014. Cham: Springer International Publishing; 2014. p. 740–55.
62. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015; Santiago, Chile: IEEE.
63. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI, USA: IEEE.
64. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL'02; 2002; USA: Association for Computational Linguistics. p. 311–8.

65. Vedantam R, Zitnick CL, Parikh D. CIDEr: consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; Boston, MA, USA; p. 4566–75.
66. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein J, Lavie A, Lin CY, Voss C, editors. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: Association for Computational Linguistics; 2005. p. 65–72.
67. Anderson P, Fernando B, Johnson M, Gould S. SPICE: semantic propositional image caption evaluation. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision–ECCV 2016. Cham: Springer International Publishing; 2016. p. 382–98.
68. Agrawal H, Desai K, Wang Y, Chen X, Jain R, Johnson M, et al. Nocaps: novel object captioning at scale. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019; Seoul, Republic of Korea: IEEE.
69. Dou ZY, Kamath A, Gan Z, Zhang P, Wang J, Li L, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22; 2024; Red Hook, NY, USA: Curran Associates Inc.
70. Hudson DA, Manning CD. GQA: a new dataset for real-world visual reasoning and compositional question answering. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, California, CA, USA: IEEE.
71. Yang J, Duan J, Tran S, Xu Y, Chanda S, Chen L, et al. Vision-language pre-training with triple contrastive learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA: IEEE.
72. Li Y, Hsiao JH, Ho C. Object prior embedded network for query-agnostic image retrieval. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2022; New Orleans, LA, USA; p. 4965–70.
73. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009; Miami, FL, USA: IEEE.
74. Yang Z, Gan Z, Wang J, Hu X, Lu Y, Liu Z, et al. An empirical study of GPT-3 for few-shot knowledge-based VQA. Proc AAAI Conf Artif Intell. 2022 Jun;36(3):3081–9. doi:10.1609/aaai.v36i3.20215.
75. Marino K, Rastegari M, Farhadi A, Mottaghi R. OK-VQA: a visual question answering benchmark requiring external knowledge. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA: IEEE.
76. Hou Z, Sun F, Chen YK, Xie Y, Kung SY. MILAN: masked image pretraining on language assisted representation. arXiv:220806049. 2022.
77. You H, Zhou L, Xiao B, Codella N, Cheng Y, Xu R, et al. Learning visual representation from modality-shared contrastive language-image pre-training. Gewerbestrasse, Cham, Switzerland: Springer Nature Switzerland; 2022. p. 69–87.
78. Deng X, Shi H, Huang R, Li C, Xu H, Han J, et al. GrowCLIP: data-aware automatic model growing for large-scale contrastive language-image pre-training. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023; Paris, France: IEEE; p. 22121–32.
79. Gupta A, Dollar P, Girshick R. LVIS: a dataset for large vocabulary instance segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA: IEEE.
80. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22; 2024; Red Hook, NY, USA: Curran Associates Inc.
81. Lu P, Mishra S, Xia T, Qiu L, Chang KW, Zhu SC, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in neural information processing systems. New Orleans, LA, USA: Curran Associates, Inc.; 2022. Vol. 35, p. 2507–21.

82. Gurari D, Li Q, Lin C, Zhao Y, Guo A, Stangl A, et al. VizWiz-Priv: a dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA: IEEE.
83. Singh A, Natarajan V, Shah M, Jiang Y, Chen X, Batra D, et al. Towards VQA models that can read. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA: IEEE.
84. Birhane A, Dehdashtian S, Prabhu V, Boddeti V. The dark side of dataset scaling: evaluating racial classification in multimodal models. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT '24; 2024; New York, NY, USA: Association for Computing Machinery; p. 1229–44.
85. Chen Q, Hong Y. MedBLIP: bootstrapping language-image pretraining from 3D medical images and texts. In: Cho M, Laptev I, Tran D, Yao A, Zha H, editors. Computer vision–ACCV 2024. Singapore: Springer Nature Singapore; 2025. p. 98–113.
86. Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021; Cambridge, MA, USA: PMLR; vol. 139, p. 4904–16.
87. Chen Z, Liu G, Zhang BW, Yang Q, Wu L. AltCLIP: altering the language encoder in CLIP for extended language capabilities. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the association for computational linguistics: ACL 2023. Toronto, ON, Canada: Association for Computational Linguistics; 2023. p. 8666–82.
88. Ko B, Gu G. Large-scale bilingual language-image contrastive learning. arXiv:220314463. 2022.
89. Wu B, Cheng R, Zhang P, Gao T, Vajda P, Gonzalez JE. Data efficient language-supervised zero-shot recognition with optimal transport distillation. arXiv:211209445. 2021.
90. Li Y, Liang F, Zhao L, Cui Y, Ouyang W, Shao J, et al. Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm. arXiv:211005208. 2021.
91. Fan L, Krishnan D, Isola P, Katabi D, Tian Y. Improving CLIP training with language rewrites. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23; 2024; Red Hook, NY, USA: Curran Associates Inc.
92. Yang K, Deng J, An X, Li J, Feng Z, Guo J, et al. ALIP: adaptive language-image pre-training with synthetic caption. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023; Paris, France; p. 2910–9.
93. Hu X, Liu D, Li H, Huang X, Shao J. VLSBench: unveiling visual leakage in multimodal safety. arXiv:241119939. 2024.
94. Zhang Y, Chen L, Zheng G, Gao Y, Zheng R, Fu J, et al. SPA-VL: a comprehensive safety preference alignment dataset for vision language model. arXiv:240612030. 2024.
95. Li L, Xie Z, Li M, Chen S, Wang P, Chen L, et al. VLFeedback: a large-scale AI feedback dataset for large vision-language models alignment. arXiv:241009421. 2024.
96. Zhou K, Yang J, Loy CC, Liu Z. Conditional prompt learning for vision-language models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA: IEEE.
97. Du Y, Wei F, Zhang Z, Shi M, Gao Y, Li G. Learning to prompt for open-vocabulary object detection with vision-language model. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA: IEEE.
98. Huang H, Chu J, Wei F. Unsupervised prompt learning for vision-language models. 2022. doi: 10.48550/arXiv.2204.03649.
99. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2020. vol. 33, p. 1877–901.
100. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT'21; 2021; New York, NY, USA: ACM.
101. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Von Arx S, et al. On the opportunities and risks of foundation models. arXiv:210807258. 2021.

102. Ghosh A, Acharya A, Saha S, Jain V, Chadha A. Exploring the frontier of vision-language models: a survey of current methodologies and future directions. arXiv:240407214. 2024.

103. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for datasets. Commun ACM. 2021 Nov;64(12):86–92. doi:10.1145/3458723.