Current State-of-art of Visual Language Models

Andrea turchet
University of Udine
Italy, Udine
152998@spes.uniud.it

ABSTRACT

Current visual recognition heavily relies on deep neural networks (DNNs) trained on vast amounts of manually annotated data. This approach, while effective, suffers from two major drawbacks: the laborious and time-consuming nature of data labeling and the need to train separate DNNs for each specific task. Vision-Language Models (VLMs) present an effective alternative by utilizing the vast amount of readily available image-text pairs on the internet. This work primarily focuses on mapping images to language and it also explores the potential extension of Vision-Language Models (VLMs) to video data.

1 INTRODUCTION

Recent years have seen major advancements in language modeling, with the emergence of Large Language Models (LLMs) such as Llama and GPT capable of addressing a diverse range of tasks. The increasing prevalence of these models underscores their growing significance; while initially limited to textual inputs, these models now incorporate visual information, opening a new era of AI applications.

Despite substantial progress in integrating language and vision, challenges persist; existing Vision Language Models (VLMs) often exhibit limitations in comprehending spatial relationships and performing counting operations without relying on complex engineering solutions and additional annotated data. Furthermore, many VLMs demonstrate a lack of understanding of attributes and order, frequently omitting portions of input prompts and necessitating intricate prompt engineering to achieve desired outcomes. The propensity for hallucination and generation of irrelevant content further underscores the ongoing need for research and development in this area. This research provides a comprehensive overview and analysis of Vision-Language Model (VLM) methodologies. It systematically categorizes and examines VLMs across three key dimensions: pre-training strategies for robust vision-language representation learning, training process and future direction of the VLMs research. It offers also a comparative analysis of existing methods, highlighting their strengths, weaknesses, and performance benchmarks.

2 INGREDIENTS OF A VLM

Visual Language Models (VLMs) represent a significant advancement in the field of artificial intelligence, combining the power of computer vision and natural language processing. These models are characterized by their diverse architectural paradigms, which vary depending on how they approach the fusion of visual and textual modalities. Despite this diversity, VLMs share a fundamental principle: the transformation of both visual and textual inputs into a unified representational format. At the core of VLMs lie embeddings

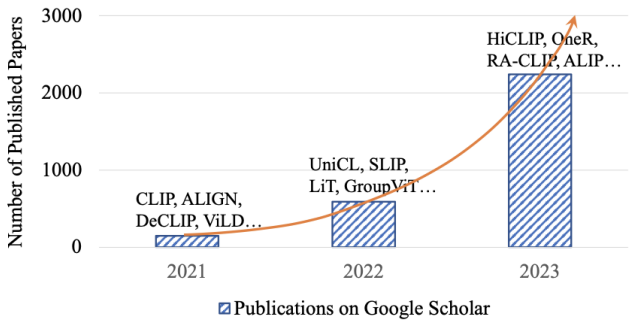

Figure 1: Number of publications on visual recognition VLMs (from Google Scholar). The publications grow exponentially since the pioneer study CLIP [10] in 2021.

- numerical vectors that encapsulate semantic information. These embeddings are fundamental in aligning and integrating visual and textual data within a shared latent space. This integration allows VLMs to understand and process information from both modalities simultaneously, enabling more sophisticated analysis and generation tasks. An aspect to be noted of VLM architecture is the timing and strategy of modality fusion. This can range from early-stage alignment, where visual and textual information is combined at the beginning of the processing pipeline, to late-stage integration, where the fusion occurs closer to the output. The choice of fusion strategy significantly impacts the model's performance and capabilities, influencing how it interprets and generates multimodal content. The architectural decisions in VLM design involve careful consideration of how to best use the strengths of both visual and textual processing while mitigating their individual limitations. As research in this field progresses, it is possible to expect to see increasingly sophisticated VLM architectures that push the boundaries of multimodal understanding and generation, opening up new possibilities in areas such as image captioning, visual question answering, and multimodal content creation.

3 FAMILIES OF VLMS

Visual recognition-related Vision-Language Model (VLM) studies have demonstrated significant advancements since the introduction of CLIP [34]. The evolution of pre-training objectives has progressed from a singular focus to a more comprehensive approach incorporating multiple hybrid objectives. Early VLMs typically employed a single pre-training objective, now contemporary models have introduced multiple objectives, including contrastive, alignment, and generative approaches. This multifaceted strategy aims to explore

the potential synergies between different objectives, ultimately leading to more robust VLMs and enhanced performance in downstream tasks. Several key paradigms have emerged in VLM pre-training. Contrastive training, a widely adopted strategy, utilizes pairs of positive and negative examples to train the VLM in predicting similar representations for positive pairs while differentiating negative pairs. The masking approach focuses on reconstructing masked image patches given unmasked text, or vice versa, reconstructing masked words in captions given unmasked images. VLMs based on pretrained backbones often employ open-source Language Models (LLMs) like Llama [43] to learn a mapping between an image encoder and the LLM, a method that is typically less computationally intensive than training text and image encoders from scratch. Generative VLMs, while often the most computationally expensive to train, are designed to generate both images and captions. It is important to note that these paradigms are not mutually exclusive; many contemporary approaches employ a combination of contrastive, masking, and generative criteria to achieve optimal performance. Pre-training frameworks have shifted from using separate networks to a unified network architecture; initial VLM designs relied on two-tower pre-training frameworks, while recent iterations have explored one-tower pre-training frameworks. These unified networks encode both images and texts within a single structure, resulting in reduced GPU memory usage and facilitating more efficient communication across data modalities. This evolution in architecture design reflects the ongoing efforts to improve the efficiency and effectiveness of VLMs in visual recognition tasks.

3.1 Contrastive-based VLMs

Contrastive learning has emerged as a prominent and solid paradigm in the pre-training of Visual Language Models (VLMs), working good in enhancing the models' ability to learn discriminative image-text features. This approach has gained significant attraction in recent years, with numerous studies demonstrating its effectiveness in improving the performance and generalization capabilities of VLMs across various multimodal tasks [34], [30], [26]. Contrastive Language-Image Pre-training (CLIP), introduced by Radford et al. in 2021, exemplifies a notable contrastive method utilizing the InfoNCE loss. CLIP's approach to defining example pairs is distinctive: it considers an image paired with its authentic caption as a positive example, while treating the same image combined with captions describing other images within the mini-batch as negative examples. CLIP's innovation lies in its ability to unify visual and linguistic information within a shared representational framework. The model achieves this by training vision and text encoders from scratch, guiding them to project images and their corresponding captions into similar embedding vectors through the application of a contrastive loss function. The original CLIP model, trained on an extensive dataset of 400 million image-caption pairs sourced from the internet, demonstrated remarkable capabilities in zero-shot classification transfer. Notably, a CLIP model utilizing a ResNet-101 architecture matched the performance of a supervised ResNet [15] model, achieving a 76.2% accuracy in zero-shot classification tasks. Furthermore, CLIP outperformed traditional supervised models on several robustness benchmarks, highlighting its adaptability to diverse visual recognition challenges. Building upon CLIP's foundation, SigLIP, proposed by [58] in 2023, introduces a key modification to the contrastive learning approach. While maintaining a similar overall structure to CLIP, SigLIP employs the original Noise Contrastive Estimation (NCE) loss, which is based on binary cross-entropy, instead of CLIP's multi-class objective derived from InfoNCE. This alteration enables SigLIP to achieve superior zero-shot performance, particularly when working with smaller batch sizes compared to CLIP. The enhanced efficiency in smaller batch scenarios makes SigLIP an attractive option for researchers and practitioners working with limited computational resources. Further advancing the field, Latent Language Image Pretraining (Llip) [23], addresses a fundamental aspect of image-text relationships: the inherent diversity in image captioning. Llip's approach involves conditioning the encoding of an image on the target caption through the implementation of a cross-attention module; this methodology acknowledges and allows for multiple accurate descriptions of a single image. By accounting for caption diversity, Llip enhances the expressivity of the learned representations andt-This increased expressivity translates into improved performance across various downstream tasks, particularly in zero-shot transfer scenarios for both classification and retrieval tasks. The Llip approach represents a significant step towards more refined and context-aware visual-language models, capable of capturing the subtleties and variations inherent in human descriptions of visual content. The evolution from CLIP to SigLIP and Llip illustrates the rapid progress in contrastive learning methods for visual-language models. Each iteration brings refinements and novel insights, pushing the boundaries of what's possible in zero-shot learning and transfer capabilities. These advancements not only improve model performance but also enhance their applicability across a wider range of real-world scenarios, from content retrieval systems to automated image analysis in various domains. Contrastive objectives aim to bring together semantically similar instances (positive pairs) while pushing apart dissimilar ones (negative pairs) in the embedding space. This approach produces the learning of discriminative visual and language features, which are very important for robust zero-shot predictions. However, contrastive learning faces two primary challenges: (1) the intricate optimization of both positive and negative pairs and (2) the reliance on a heuristic temperature hyperparameter to control feature discriminability.

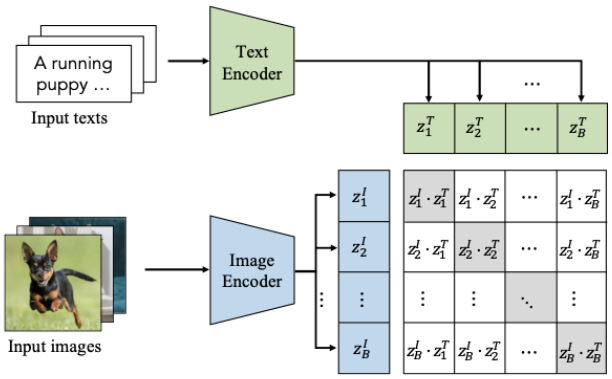

Figure 2: Illustration of the image-text contrastive learning in CLIP [34]. Figure is reproduced from [34]

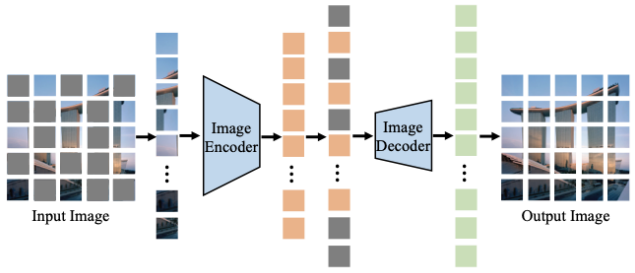

Figure 3: Illustration of masked image modelling. Figure is reproduced from [14]

3.2 VLMs with masking objectives

Masking, a core technique in deep learning, involves selectively obscuring parts of input data. Rooted in denoising autoencoders [47] and related to image inpainting [31], masking has been employed to learn robust representations. The BERT model [9] popularized Masked Language Modeling (MLM) for natural language processing, while Masked Image Modeling (MIM) techniques like MAE [14] and I-JEPA [2] emerged in computer vision. Given its compatibility with tokenization, masking is particularly effective with transformer architectures [46]. Several works have combined MLM and MIM for Vision-Language Models (VLMs). FLAVA [37] incorporates masking as part of its training strategy, while MaskVLM [22] is dedicated to this approach.

3.2.1 FLAVA. The model comprises three transformer-based encoders: one for images, one for text, and a multimodal encoder that fuses image and text representations. During pretraining, the model is trained on both paired and unpaired image-text data. For paired data, contrastive, masked multimodal modeling (MMM), and image-text matching (ITM) losses are employed. For unpaired data, masked image modeling (MIM) and masked language modeling (MLM) losses are applied to the respective encoders. This comprehensive training regime allows FLAVA to excel at a wide range of tasks, including image and text understanding as well as multimodal reasoning.

3.3 Generative-based VLMs

Unlike traditional methods that focus on creating abstract representations of images and text and then correlating them, the generative approach directly handles the creation of text and/or images themselves. Some methods, like CoCa [54], develop a complete text encoder and decoder that enables image captioning. Others, such as Chameleon Team [41] and CM3leon [56], are multimodal generative models explicitly trained to generate both text and images. Finally, there are models designed specifically to generate images from text, including Stable Diffusion [4], Imagen [17], and Parti [55]. Despite being trained only for image generation, these models can also be adapted to address various vision-language understanding tasks. This approach encourage VLMs to develop deep understandings of visual, textual, and multimodal contexts; to enhance their ability to perform zero-shot tasks, these generative objectives are often employed in conjunction with other pre-training techniques. While generative classifiers demand greater computational resources during inference, they offer substantial advantages; they exhibit an high level of robustness, demonstrating superior performance on out-of-distribution data compared to discriminative models while maintaining comparable in-distribution accuracy. Generative classifiers excel in tasks requiring compositional reasoning, such as Winoground [42], surpassing discriminative counterparts. Their alignment with human judgment [19] in terms of shape bias is another strength. Moreover, generative models can be complementarily combined with discriminative ones at test time using unlabeled data, leading to performance improvements across various tasks, especially in dynamic environments [32].

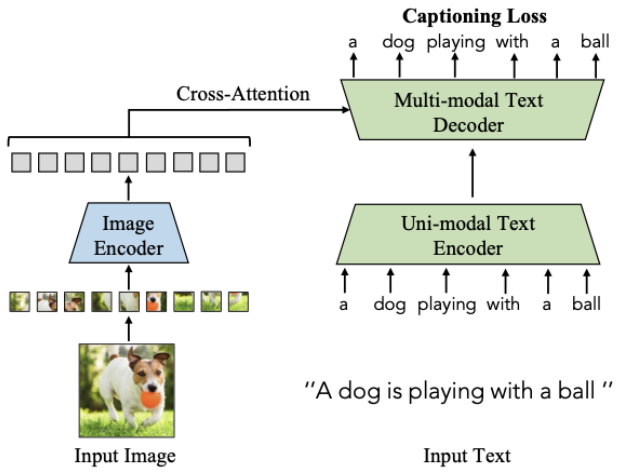

Figure 4: A simplified illustration of image-to-caption generation in COCA [54]. Figure is reproduced based on [54].

3.4 VLMs from Pretrained Backbones

Training Vision-Language Models (VLMs) from scratch is computationally expensive, demanding large amounts of data and computational resources. Consequently, researchers have focused on employing pre-trained language models and visual feature extractors to reduce training costs. The availability of open-source large language models (LLMs) has been an essential element for this approach. By learning a mapping between text and image modalities, these models can acquire the ability to answer visual questions with significantly less computational overhead. This section highlights two representative models: Frozen [45], a pioneering work in this area, and the Mini-GPT family [62]. For example, MiniGPT-5 [59] builds upon MiniGPT-4 by enabling outputs that combine text and images. To facilitate image generation, MiniGPT-5 uses generative tokens—special visual tokens that are transformed into feature vectors via transformer layers; these vectors are then fed into a frozen Stable Diffusion 2.1 model. The authors used supervised training on downstream tasks, like a multimodal dialogue generation and story generation.

In conclusion, VLM pre-training models the relationship between vision and language through various objectives, including contrastive learning, masked modality modeling, image-to-text generation, and image-text/region-word matching. To take advantage of the potential of each modality, models often incorporate single-modal tasks like masked image or language modeling. While recent work has focused on capturing global vision-language correlations for image-level tasks, other research has explored modeling fine-grained relationships between image regions and words to enhance object detection and semantic segmentation [53] [27] [24] [10] [35].

4 VLM TRAINING

Recent research emphasized the important role of scale in enhancing deep neural network performance [16]. As consequence, many studies have prioritized increasing computational resources and dataset size to develop superior models; this trend is exemplified by CLIP, trained on a massive dataset of 400 million images using substantial compute power. Even its open-source counterpart, OpenCLIP [18], required 256-600 GPUs for training. However, a groundbreaking study [38] challenges the notion that scale alone determines model performance, proving that meticulously curated datasets can surpass models trained with significantly higher computational power. About the training data, DataComp benchmark evaluates the impact of dataset quality on VLM performance by comparing different filtering strategies on large-scale web data. The findings underscore that the practice of data pruning is a critical component in the training of Vision-Language Models; keep note that data-pruning methods can be categorized into heuristic, bootstrapping, and diversity-based approaches.

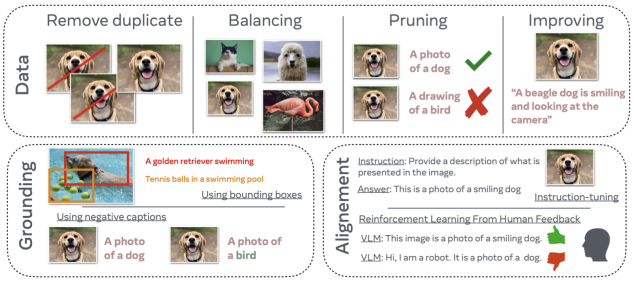

Figure 5: Data quality and quantity are one of the most important aspects of VLM training. Diverse, balanced, and clean datasets are essential for robust models. Effective data curation involves removing duplicates, filtering irrelevant content, and enhancing caption quality. Grounding ensures accurate word-to-image associations, while alignment brings model outputs into line with human expectations.

5 SOFTWARE AND RESOURCES FOR VLM TRAINING

This section explores existing software and resources for VLM evaluation and training.

5.1 Public Software Repositories

Fortunately, numerous open-source software repositories exist to empower VLM exploration. Tools like OpenCLIP and Transformers implement various VLM architectures; these platforms are ideal for benchmarking and comparing pre-trained VLMs on specific downstream tasks.

5.2 Understanding Compute Resource Needs

The computational resources required for VLM training significantly impact associated costs. Models like CLIP utilized massive compute power exceeding 500 GPUs, translating to hundreds of thousands of dollars – often out of reach for most research labs or companies. However, by employing high-quality datasets and the use of masking strategies for larger models, training a contrastive model like CLIP from scratch on vast image datasets (hundreds of millions) might not require more than 64 GPUs (roughly equivalent to $10,000 in compute costs). Additionally, if pre-trained image or

text encoders or LLMs are used as the foundation, the learning cost for mapping should be substantially lower.

5.3 Accelerating Training

Recent software advancements like PyTorch's torch.compile significantly speed up model training. The xformers library with more efficient attention mechanisms also contributes to faster training times. However, data loading is often overlooked as a major training bottleneck, especially when dealing with large image batches. This is because large-scale datasets are often stored in compressed chunks that need on-the-fly decompression, slowing down training. Here are some key recommendations for faster training:

- **Store uncompressed files:** Optimize training speed by storing as many uncompressed data files as possible.
- **FFCV library:** Fast Forward Computer Vision (FFCV) library allow to create data files that offer significantly faster loading compared to webdataset.
- **Masking:** Masking image tokens during training can considerably boost training efficiency while improving model performance, especially for models with billions or hundreds of millions of parameters [25].

While these approaches may require more storage space due to uncompressed files, the speed benefits often outweigh the cost; the reduced compute time translates to lower overall training costs.

5.4 Importance of Hyperparameters

A recent study by [28] explored the most used design choices for VLM training. Their findings highlight image resolution, visual encoder capacity, and visual pre-training data as the most important factors impacting model performance. Interestingly, the study suggests that the method of modality connection has a lesser influence. The authors also study the importance of diverse training data, ranging from text-only to interleaved and image-caption paired data. They demonstrate that the optimal mix of these data types achieves the best performance across zero-shot classification and visual question answering tasks.

6 WHAT MODEL TO USE?

There are several methods to train VLMs; some of them use simple contrastive training criteria, others use masking strategies to predict missing texts or image patches, while some models are using generative paradigms such as autoregression or diffusion. It is also possible to employ a pre-trained vision or text backbones like Llama or GPT. In that instance, building a VLM model requires learning only a mapping between the LLM and vision encoder representations. So, from all those methods, which one should someone choose? Do we need to train vision and text encoder from scratch like CLIP or is it better to start with pretrained LLM such as Flamingo or MiniGPT?

6.1 When to use contrastive models like CLIP?

Contrastive models like CLIP associate text with visual concepts while keeping a simple training paradigm by pushing text and image representation to be matched in the representation space. By doing so, CLIP learns representations that have both meaning in the image and text space, which makes it possible to prompt the CLIP text encoder with words such that we can retrieve the images that map to the corresponding text representations. For example, many data curation pipelines such as MetaCLIP [50] are using metadata string matching to build datasets to ensure that each word or concept has enough images associated with them. CLIP models are also a good base for building more complex models, especially when trying to improve grounding. For researchers who are looking at trying additional training criteria or different model architectures to better capture relations or a better understanding of concepts, CLIP is a particularly good starting point. However, one should keep in mind that CLIP is not a generative model, thus it is not possible to generate a caption given a specific image. It is only possible to retrieve the best caption within a list of already existing captions. In consequence, current CLIP models cannot be used to provide high-level descriptions of a given image. Another drawback is that CLIP usually needs a very large dataset as well as large batch sizes to offer decent performances, which implies that CLIP usually needs significant resources to be trained from scratch.

6.2 When to use masking?

Masking is an alternative strategy to train VLMs. By learning to reconstruct data from both masked images and text, it is possible to jointly model their distributions. On the other hand, contrastive models which operate in a representation space, models based on masking might need to employ a decoder to map back the representation to the input space (and thus to apply a reconstruction loss). Training an additional decoder might add an additional bottleneck which might make these methods less efficient than a purely contrastive one. However, the advantage is that there is no batch dependency anymore since each example can be considered separately (because there is no need negative examples). Removing negative examples can enable the use of smaller mini-batches without the need to finetune additional hyper-parameters such as the softmax temperature. Many VLM methods use a mix of masking strategies along with some contrastive loss.

6.3 When to use a generative model?

Generative models based on diffusion or autoregressive criteria have demonstrated impressive abilities in generating photorealistic images based on text prompt. Most large-scale training efforts on VLM are also starting to integrate image generation components. Some researchers explain that having the ability to generate images given words is an important step towards creating a good world model while other researchers argue that such a reconstruction step is not needed [3]. However from an application perspective, it might be easier to understand and assess what the model has learned when it is able to decode abstract representations in the input data space (it is also possible to add a decoder on top of a trained join-embedding architecture [5]). While models like CLIP would need extensive k-NN evaluations using millions of image data points to show what the images closest to a given word embedding look like, generative models can just output the most probable image directly without such an expensive pipeline. In addition, generative models can learn an implicit joint distribution between text and images which might be more suited for learning good representations than employing pretrained unimodal encoders. However, they are more

computationally expensive to train than their contrastive learning counterpart.

6.4 When to use LLM on pretrained backbone?

Using already pretrained text or vision encoder can be a good alternative when having access to limited resources. In that case, only the mapping between the text representation and vision representation should be learned. However, the main issue with this approach is that the VLM will be impacted by the potential hallucination of the LLM. It could also be impacted by any bias coming from the pretrained models. In consequence, there might be an additional overhead in trying to correct the defect of the vision model or of the LLM. Some might argue that it is important to employ independent image and text encoder to project the information into a lower dimension manifold on which it is possible to learn a mapping while others might argue that it is important to learn the distribution of image and text jointly. To summarize, using a pretrained model is interesting when having limited access to compute resources and when researchers are interested in learning mapping in representation spaces.

7 VLM AND VIDEOS

Video data presents new challenges and opportunities for models, including understanding motion and dynamics, localizing objects and actions in space and time, and generating text-to-video content. Recent years have seen the emergence of fundamental computer vision tasks such as text-to-video retrieval, video question answering, and generation [50][40][6]. Video data, however, requires significant computational resources and storage, particularly when dealing with high-frame-rate videos. This necessitates trade-offs in VLMs, such as using compressed video formats (e.g., H.264 encoding) with on-the-fly video decoders, initializing video encoders from image encoders, and incorporating spatial and temporal pooling/masking mechanisms [11][13]. Non-end-to-end VLMs, which extract video features offline and train models on these features, are also being explored. Early video-text models trained from scratch the visual and text components using self-supervised criteria [1]. However, unlike image models, contrastive video-text models were not the primary approach, and early fusion and temporal alignment of video and text were preferred [39]. This is because more temporal granularity in the representation is often more interesting than computing a global representation of the video. Recently, a trend similar to image-language models has emerged for video-language models: pre-trained Large Language Models are being used and aligned with video encoders to augment the LLMs with video understanding capabilities. Modern techniques such as visual instruction tuning are also being adapted to video. The first successful general approach to video-language modeling was VideoBERT [39], which is an early fusion approach that fuses visual and textual tokens representing video caption pairs with a single transformer network. The video data comes from YouTube instructional cooking videos, and the aligned text is obtained using automatic speech recognition (ASR). VideoBERT demonstrates strong alignment and it is the first model able to perform well on video tasks that require generating text, such as zero-shot action classification and open-ended video captioning. Going beyond global video and text alignment, Multimodal Event Representation Learning Over Time (MERLOT) [57] achieves video language alignment by temporally aligning text with video. Unlike VideoBERT, which is trained on curated instructional cooking videos, MERLOT is trained on a large-scale dataset of YouTube videos with diverse content and corresponding text obtained by ASR. MERLOT uses a transformer network trained in a purely self-supervised way with a contrastive objective, masked language modeling objective, and temporal reordering objective. The model demonstrated impressive capabilities on question answering tasks, particularly visual common sense reasoning. However, it lacks the ability to generate text, which limits its ability to demonstrate advanced visual reasoning capabilities.

8 FUTURE STUDIES

VLM has achieved significant success in various visual recognition tasks, enabling effective usage of web data, zero-shot prediction without task-specific fine-tuning, and open-vocabulary visual recognition of images of arbitrary categories. However, there are several research challenges and potential directions that could be followed in the future to further improve VLMs.

8.1 Pre-training

- **Fine-grained Vision-Language Correlation Modeling:** VLMs can benefit from fine-grained vision-language correspondence knowledge to recognize patches and pixels beyond images, which is crucial for dense prediction tasks such as object detection and semantic segmentation. More research is needed in this direction [53] [27] [24] [10] [51] [60].
- **Unification of Vision and Language Learning:** Unifying image and text learning within a single Transformer can enable efficient communication across data modalities, improving training effectiveness and efficiency. This issue has attracted some attention, but more efforts are needed [44] [20].
- **Pre-training VLMs with Multiple Languages:** Most existing VLMs are trained with a single language, which can introduce bias and hinder VLM applications in other language areas [36] [7]. Pre-training VLMs with texts of multiple languages can enable VLMs to work efficiently and effectively across different language scenarios [8] [21].
- **Data-Efficient VLMs:** Training VLMs with limited image-text data can mitigate sustainability concerns. More research is needed on data-efficient VLMs, such as learning from supervision among image-text pairs [49] [26].
- **Pre-training VLMs with Large Language Models (LLMs):** Recent studies [12] [52] have explored using LLMs to enhance VLM pre-training by retrieving rich language knowledge. More research is expected in this direction.

8.2 Transfer Learning

- **Unsupervised VLM Transfer:** Most existing VLM transfer studies work with supervised or few-shot supervised setups, which can lead to overfitting. Unsupervised VLM transfer allows exploring massive unlabelled data with lower risk of overfitting.

- **VLM Transfer with Visual Prompt/Adapter:** Most existing studies [61] focus on text prompt learning, while visual prompt learning or visual adapter can enable pixel-level adaptation in various dense prediction tasks.
- **Test-time VLM Transfer:** Most existing studies conduct transfer by fine-tuning VLMs on each downstream task, which can be repetitive. Test-time VLM transfer allows adapting prompts on the fly during inference, avoiding repetitive training.
- **VLM Transfer with LLMs:** Several attempts [33][29] have explored using LLMs to generate text prompts that better describe downstream tasks. More research is expected in this direction.

8.3 VLM Knowledge Distillation:

VLM knowledge distillation can be further explored from two aspects: knowledge distillation from multiple VLMs and knowledge distillation for other visual recognition tasks such as instance segmentation, panoptic segmentation, and person reidentification.[48].

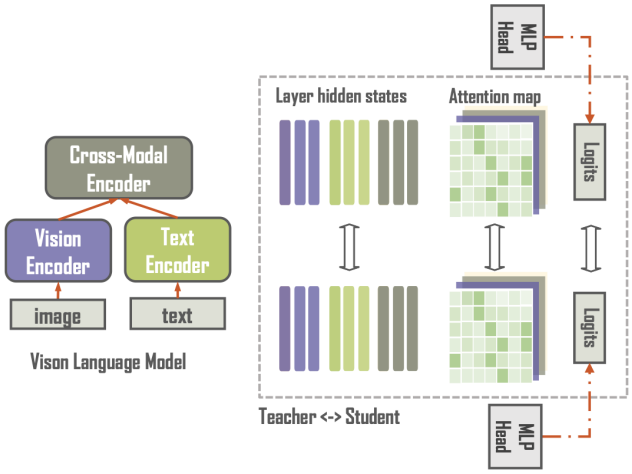

Figure 6: pre-training with knowledge distillation [48]

9 CONCLUSION

This research has demonstrated the diverse techniques available for efficient VLM development, each with its own strengths and weaknesses. The optimal approach depends on specific application requirements and available resources. As the field continues to rapidly evolve, this research provides a comprehensive overview of the current state of VLM pre-training, offering a reading for understanding the latest advancements and future directions.

REFERENCES

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4575–4583.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15619–15629.

[3] Randall Balestriero and Yann LeCun. 2024. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337* (2024).

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[5] Florian Bordes, Randall Balestriero, and Pascal Vincent. 2021. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164* (2021).

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators.

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3558–3568.

[8] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679* (2022).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems* 35 (2022), 32942–32956.

[11] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6824–6835.

[12] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems* 36 (2024).

[13] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems* 35 (2022), 35946–35958.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701* (2020).

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

[18] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773

[19] Priyank Jaini, Kevin Clark, and Robert Geirhos. 2023. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779* (2023).

[20] Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. 2023. Unifying vision-language representation space with single-tower transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 980–988.

[21] Byungsoo Ko and Geonmo Gu. 2022. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463* (2022).

[22] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131* (2022).

[23] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wildon, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740* (2024).

[24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.

[25] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23390–23400.

[26] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208* (2021).

[27] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*. PMLR, 23033–23044.

[28] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611* (2024).

[29] Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183* (2022).

[30] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*. Springer, 529–544.

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2536–2544.

[32] Mihir Prabhudesai, Tsung-Wei Ke, Alex Li, Deepak Pathak, and Katerina Fragkiadaki. 2024. Test-time adaptation of discriminative models via diffusion generative feedback. *Advances in Neural Information Processing Systems* 36 (2024).

[33] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15691–15701.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Marina Meila and Tong Zhang (Eds.), Vol. 139. PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[35] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander T Toshev, and Jonathon Shlens. 2022. Perceptual grouping in vision-language models. (2022).

[36] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*. 2556–2565.

[37] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15638–15650.

[38] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=UmvSlP-PyV

[39] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7464–7473.

[40] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.

[41] Chameleon Team. [n.d.]. Chameleon: Mixed-modal early-fusion foundation models. CoRR, abs/2405.09818, 2024. doi: 10.48550. *arXiv preprint ARXIV.2405.09818* ([n. d.]).

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

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[44] Michael Tschannen, Basil Mustafa, and Neil Houlsby. 2022. Image-and-language understanding from pixels only. *arXiv preprint arXiv:2212.08045* 3 (2022).

[45] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[47] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. 1096–1103.

[48] Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. 2022. EfficientVLM: Fast and Accurate Vision-Language Models via Knowledge Distillation and Modal-adaptive Pruning. arXiv:2210.07795 [cs.CL] https://arxiv.org/abs/2210.07795

[49] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Peter Vajda, and Joseph E Gonzalez. 2021. Data efficient language-supervised zero-shot recognition with optimal transport distillation. *arXiv preprint arXiv:2112.09445* (2021).

[50] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data. *arXiv preprint arXiv:2309.16671* (2023).

[51] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18134–18144.

[52] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2922–2931.

[53] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. 2022. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems* 35 (2022), 9125–9138.

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

[55] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.

[56] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591* 2, 3 (2023).

[57] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems* 34 (2021), 23634–23651.

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11975–11986.

[59] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239* (2023).

[60] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16793–16803.

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.

[62] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).