Available Online: 10 June 2025 DOI: 10.54254/2977-3903/2025.23982

A survey on pre-training and transfer learning for multimodal Vision-Language Models

Zhongren Liang

Beijing University of Posts and Telecommunications, Beijing, China

2936721262@qq.com

Abstract. In recent years, Vision-Language Models (VLMs) have emerged as a significant breakthrough in multimodal learning, demonstrating remarkable progress in tasks such as image-text alignment, image generation, and semantic reasoning. This paper systematically reviews current VLM pretraining methodologies, including contrastive learning and generative paradigms, while providing an in-depth analysis of efficient transfer learning strategies such as prompt tuning, LoRA, and adapter modules. Through representative models like CLIP, BLIP, and GIT, we examine their practical applications in visual grounding, image-text retrieval, visual question answering, affective computing, and embodied AI. Furthermore, we identify persistent challenges in fine-grained semantic modeling, cross-modal reasoning, and cross-lingual transfer. Finally, we envision future trends in unified architectures, multimodal reinforcement learning, and domain adaptation, aiming to provide systematic reference and technical insights for subsequent research.

Keywords: Vision-Language Models, multimodal learning, pre-training, transfer learning, contrastive learning

1. Introduction

Recent years have witnessed multimodal learning becoming a pivotal research direction in artificial intelligence, particularly in the fusion of visual and linguistic modalities. Traditional image classification and text generation models typically operate independently, struggling with cross-modal semantic understanding and reasoning tasks. With the increasing availability of large-scale image-text paired data and continuous advancements in deep learning architectures, Vision-Language Models (VLMs) represented by CLIP (Contrastive Language–Image Pretraining) and ALIGN (A Large-scale ImaGe and Noisy-text embedding) have rapidly evolved. These models enable zero-shot image classification, open-vocabulary detection, and other cross-modal tasks, driving the practical application of multimodal intelligence.

Currently, multimodal models are transitioning from perceptual understanding to higher-level cognitive reasoning, with research focus expanding from basic alignment capabilities to transfer learning, generative abilities, and reasoning capacities. Notably, recent multimodal architectures incorporating diffusion models and reinforcement learning have endowed VLMs with enhanced generative capabilities and interactive intelligence.

This paper systematically reviews the developmental trajectory of VLMs, emphasizing mainstream pretraining paradigms and transfer learning strategies, while demonstrating their effectiveness and practical value through experimental data and application case studies. Additionally, we explore emerging applications in affective computing and embodied intelligence, identifying existing challenges and potential future directions.

2. Pre-training methods

Pre-training is crucial for VLMs to acquire generalizable multimodal representations. We categorize common pre-training approaches into three main types: contrastive learning, generative modeling, and alignment-based objectives.

2.1. Contrastive learning

Contrastive learning is a core strategy in multimodal pretraining. Models like CLIP and ALIGN employ dual-encoder architectures, independently encoding images and text while maximizing the similarity of matched image-text pairs to achieve modality alignment [1].

CLIP utilizes the meticulously curated WIT dataset, comprising approximately 400 million high-quality image-text pairs, whereas ALIGN leverages automatically collected web-scale data (about 1 billion pairs) with higher noise levels [2]. CLIP's text is more standardized, while ALIGN's text resembles “wild” web data, making its generalization more reliant on data volume [3].

CLIP adopts a Transformer-based text encoder, outperforming ALIGN's lightweight BERT variant in linguistic reasoning [4].

In zero-shot image classification, CLIP exhibits slightly better performance on ImageNet [5], while ALIGN excels in image-text retrieval tasks, highlighting its stronger cross-modal representation capabilities [6]. These differences underscore the asymmetric impact of textual semantic precision and training data quality on multimodal alignment. Higher-quality text data significantly enhances generalization for complex tasks, whereas large-scale raw data better captures diverse image-text associations.

2.2. Generative modeling

Generative multimodal models often employ encoder-decoder architectures, optimizing objectives like text generation, image generation, or image-text reconstruction to learn robust cross-modal representations. These models are well-suited for tasks such as image captioning and visual question answering [7].

For instance, freezing an image encoder and fine-tuning a multimodal Transformer demonstrates efficient generalization in few-shot settings, outperforming composite models like CLIP+BLIP in VQA and image captioning tasks [8]. BLIP enhances contrastive learning with generative pretraining (e.g., caption prediction or image-text reconstruction) in a two-stage process, while GIT frames the entire training as an image-to-text generation task [9]. GIT achieves state-of-the-art results on MSCOCO Caption and NoCaps datasets [10], excelling in descriptive text generation but underperforming in discriminative tasks like retrieval [11]. This highlights the complementary strengths of generative and contrastive paradigms for different tasks.

2.3. Alignment objectives

The core challenge in pretraining is constructing a semantically consistent embedding space. Current methods fall into two categories:

Dual-encoder architectures (e.g., CLIP, ALIGN) can encode images and text independently, matching them via similarity functions (e.g., dot product). These are computationally efficient and scalable, ideal for large-scale retrieval and classification.

Fusion-encoder architectures (e.g., ViLBERT, UNITER) can capture finer-grained interactions, excelling in tasks like VQA and visual reasoning [12].

Dual-encoders perform better in zero-shot classification and open-vocabulary retrieval, while fusion models excel in tasks requiring fine-grained interactions [13]. Hybrid approaches like BLIP-2 combine their strengths: a frozen image encoder + fused language model balances efficiency and performance [14].

Recent models like BLIP-2 (Bootstrapped Language-Image Pretraining) integrate dual-encoder and fusion-encoder advantages, using a frozen visual encoder for question-guided answering and a language decoder for answer generation [15]. Lightweight solutions like multimodal adapter modules (single-modal + cross-modal + MoE routing) enable incremental modality expansion with only 2% parameter storage, outperforming traditional methods like X-InstructBLIP [16].

3. Model architectures

To enhance adaptability and training efficiency in multitask scenarios, researchers have proposed various transfer learning methods. We focus on three representative strategies: Dual-encoder Models, Fusion-encoder Models and Encoder-decoder Models.

3.1. Dual-encoder models

CLIP's zero-shot learning heavily relies on prompt engineering. Early methods used fixed templates (e.g., “a photo of a __”), but their generalization was limited [17]. Subsequent approaches like CoOp (Context Optimization) introduced learnable soft prompts, while Prompt Ensembling improved robustness by aggregating predictions from multiple templates [18]. On zero-shot ImageNet classification, CoOp outperformed fixed templates by 4–5% in accuracy.

3.2. Fusion-encoder models

To achieve efficient multitask transfer, feature adapters have become a research hotspot in recent years. Common methods include Adapter, LoRA (Low-Rank Adaptation), and Prefix Tuning [19]. Adapter inserts a small number of trainable layers into the pre-trained model and updates only these layers to adapt to new tasks; LoRA performs low-rank decomposition on the weight matrices of the model, reducing the number of parameters to be updated and improving fine-tuning efficiency [20]; Prefix Tuning adds learnable prefix tokens to the input sequence to guide the model's behavior.

Taking BLIP-Adapter as an example, under the condition of training only the Adapter module, its performance on visual question answering datasets such as VQA v2 and GQA has already approached that of fully fine-tuned models, demonstrating extremely high parameter efficiency [21].

3.3. Encoder-decoder models

To meet the demands of low-resource or edge device deployment, knowledge distillation has been widely applied to the miniaturization of multimodal models. This strategy enables the smaller model to learn the output distribution of the larger model through the distillation process, thereby achieving knowledge transfer [22].

Research represented by MiniCLIP shows that even with significant compression of model parameters, the distilled smaller model can still maintain high performance across multiple tasks, demonstrating strong practical value.

4. Typical downstream application scenarios and future challenges

4.1. Zero-shot image classification and open-vocabulary detection

On benchmarks such as ImageNet, ObjectNet, and ImageNet-R, CLIP has demonstrated superior zero-shot classification capabilities. RegionCLIP and OV-DETR use vision-language embeddings for category matching in open-vocabulary object detection, enabling detection of new categories without annotations.

Platforms like Taobao and JD.com have deployed CLIP-based product image-text retrieval systems to enable “search by image” functionality [23]. In the autonomous driving field, multimodal models are used for image-navigation instruction understanding, integrating perception and decision-making [24]. In medical diagnosis, VLMs can be applied to combined image and clinical text analysis, improving the efficiency of clinical deployment [25].

4.2. Visual semantic segmentation

Although vision-language pretraining models were not originally designed for pixel-level segmentation tasks, subsequent research such as CLIPSeg has achieved “text-guided semantic segmentation” by aligning text embeddings with image feature maps; GroupViT introduced the concept of group tokens to aggregate semantic regions within images, partially alleviating the spatial resolution limitations of CLIP models and expanding the application boundaries of multimodal models in fine-grained visual tasks [26]. However, visual semantic segmentation still faces numerous challenges: the spatial resolution bottleneck of pretrained models limits precise capture of fine-grained visual information; the high cost and scarcity of pixel-level annotations restrict model training and generalization capabilities; the accuracy of cross-modal alignment, especially in complex scenarios requiring precise correspondence between text and image regions, needs improvement; moreover, models exhibit insufficient generalization and robustness in new environments, and the computational cost is high, making it difficult to meet real-time application demands. Future work should explore multi-task collaborative frameworks to promote the integration of visual semantic segmentation with object detection, instance segmentation, and other tasks, thereby enhancing overall performance and practical value.

4.3. Image-text retrieval and matching

In the task of image-text retrieval and matching, CLIP’s dual-encoder architecture offers good scalability and computational efficiency, making it suitable for large-scale image-text pair matching scenarios. In contrast, fusion models such as ViLT and ALBEF employ cross-attention mechanisms to jointly model images and text, further enhancing fine-grained semantic matching performance, which is especially suitable for applications requiring higher semantic consistency [27]. However, image-text retrieval and matching still face numerous challenges: how to further narrow the cross-modal semantic gap between images and text to improve deep semantic understanding; how to achieve more precise fine-grained matching when handling complex, ambiguous, or lengthy texts; how to cope with retrieval efficiency and storage demands brought by large-scale data; how to handle diverse and noisy multimodal data to ensure the stability and robustness of matching; how to enhance model generalization and adaptability across domains and languages; and how to integrate user interaction and personalized recommendations to improve system intelligence and user experience. Future research needs to make continuous breakthroughs in these areas to meet the increasingly complex and diverse application demands.

4.4. Applications in embodied intelligence and affective computing

Multimodal models are being extended to the field of embodied intelligence (Embodied AI), such as enabling intelligent robots to understand spoken commands, recognize environmental images, and achieve "seeing + hearing + acting" [28].

In affective computing, modalities including vision, language, and audio are increasingly integrated to enhance comprehensive understanding of human emotions, tone, facial expressions, and other information. For example, EmotiCLIP attempts to incorporate emotion labels to assist contrastive training, strengthening the model's ability to capture emotional states [28].

Currently, VLMs still lack sufficient capacity for high-level semantic modeling such as object relationships and spatial layouts. However, recent research like Omni-Scene combines Gaussian unified representations with diffusion models to enable multimodal generation in autonomous driving scenarios. Its dual decoding mechanism for volume and pixels enhances fine-grained geometric modeling [28].

Such approaches may help overcome VLMs' shortcomings in spatial reasoning.

BEIT-3 has explored multilingual pretraining by reusing language model pipelines to achieve cross-lingual transfer. Additionally, remote sensing vision-language models (such as SkyCLIP) leverage geographic semantic label generation, demonstrating the potential for vertical domain adaptation and providing new ideas for cross-lingual and cross-domain transfer [29].

5. Conclusion

With the continuous evolution of model architectures and diversification of training strategies, vision-language models are rapidly advancing toward higher levels of cognitive intelligence. From contrastive learning to generative paradigms, from prompt engineering to parameter-efficient transfer methods such as prompt tuning and LoRA, multimodal research is gradually breaking down modality boundaries to achieve more natural human-computer interaction capabilities. Future development will focus on three core directions:

First, enhancing the model's ability to represent spatial structures and entity relationships to drive significant improvements in visual grounding tasks;

Second, strengthening the model's sensory interaction and action understanding capabilities in embodied VLMs;

Third, building multimodal foundation models with unified representations and task generalization capabilities to support cross-lingual and cross-domain knowledge transfer.

Through continuous mining of high-quality image-text data and efficient design of adaptation mechanisms, multimodal intelligence will play a broader role in cognitive reasoning, affective understanding, and real-world decision-making, becoming a key pillar in the architecture of general artificial intelligence.

References

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 8748–8763. Retrieved from https://proceedings.mlr.press/v139/radford21a.html

[2] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv preprint arXiv:2205.01917*. Retrieved from https://arxiv.org/abs/2205.01917

[3] Chen, D., Zhang, Y., Wang, Z., & Li, H. (2022). ProtoCLIP: Prototypical Contrastive Language Image Pretraining. *arXiv preprint arXiv:2206.10996*. Retrieved from https://arxiv.org/abs/2206.10996

[4] Joshi, S., Jain, A., Payani, A., & Mirzasoleiman, B. (2024). Data-Efficient Contrastive Language-Image Pretraining: Prioritizing Data Quality over Quantity. *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, *238*, 1000–1008. Retrieved from https://proceedings.mlr.press/v238/joshi24a.html

[5] Cui, Y., Zhao, L., Liang, F., Li, Y., & Shao, J. (2022). Democratizing Contrastive Language-Image Pre-training: A CLIP Benchmark of Data, Model, and Supervision. *arXiv preprint arXiv:2203.05796*. Retrieved from https://arxiv.org/abs/2203.05796

[6] Pan, X., Ye, T., Han, D., Song, S., & Huang, G. (2022). Contrastive Language-Image Pre-Training with Knowledge Graphs. *arXiv preprint arXiv:2210.08901*. Retrieved from https://arxiv.org/abs/2210.08901

[7] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *Proceedings of the 39th International Conference on Machine Learning*, *162*, 12888–12900. Retrieved from https://proceedings.mlr.press/v162/li22n.html

[8] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *Proceedings of the 40th International Conference on Machine Learning, 202*, 19730–19742. Retrieved from https://proceedings.mlr.press/v202/li23q.html

[9] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., & Wang, L. (2022). GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*. Retrieved from https://arxiv.org/abs/2205.14100

[10] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hassani, A., Jeong, J., Sezer, U., Alabdulmohsin, I., Smaira, L., Raposo, D., Tyszkiewicz, M., et al. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198.* Retrieved from https://arxiv.org/abs/2204.14198

[11] Liu, Y., Zhang, Y., Wang, Y., Hou, L., Cao, J., & Bao, J. (2023). BEIT-3: Scaling Multimodal Transformers Across Vision, Language, and Audio. *arXiv preprint arXiv:2302.00915*. Retrieved from https://arxiv.org/abs/2302.00915

[12] Jia, C., Yang, Y., Xia, Y., Chen, K., Parekh, Z., Pham, H., ... & Zettlemoyer, L. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML).* https://arxiv.org/abs/2102.05918

[13] Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *International Conference on Machine Learning (ICML).* https://arxiv.org/abs/2102.03334

[14] Li, J., Zhu, Y., Zhang, Y., Yin, X., Lu, J., & Li, X. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS 2023)*. https://arxiv.org/abs/2301.12597

[15] Pfeiffer, J., Rücklé, A., Dürr, J., Frank, A., & Gurevych, I. (2021). AdapterFusion: Non-Destructive Task Composition for Transfer Learning. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL).* https://arxiv.org/abs/2005.00247

[16] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From *Natural Language Supervision. arXiv preprint*. https://arxiv.org/abs/2103.00020

[17] Sun, H., Wang, Y., &Xu, L. (2025). Parrot: Multilingual Visual Instruction Tuning. *arXiv preprint arXiv:2406.02539*. Retrieved from https://arxiv.org/abs/2406.02539

[18] Lai, W., Mesgar, M., & Fraser, A. (2025). LLMs Beyond English: Scaling Multilingual Capability with Cross-Lingual Feedback. *arXiv preprint arXiv:2406.02540.* Retrieved from https://arxiv.org/abs/2406.02540

[19] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Wang, Y., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685.* Retrieved from https://arxiv.org/abs/2106.09685

[20] Zheng, Y., Lin, K., Wang, J., et al. (2025). PlanAgent: A Multi-modal Large Language Agent for Closed-loop Vehicle Motion Planning. *arXiv preprint arXiv:2406.01587.* Retrieved from https://arxiv.org/abs/2406.01587

[21] Zhou, H., Li, M., Zhang, F., et al. (2025). UniQA: Unified Vision-Language Pre-training for Image Quality and Aesthetic Assessment. *arXiv preprint arXiv:2406.01069.* Retrieved from https://arxiv.org/abs/2406.01069

[22] Wang, H., Dong, K., Zhu, Z., et al. (2024). Transferable Multimodal Attack on Vision-Language Pre-training Models. *Proceedings of the IEEE Symposium on Security and Privacy*. https://doi.org/10.1109/sp54263.2024.00102

[23] Zhang X ,Guo C .Research on Multimodal Prediction of E-Commerce Customer Satisfaction Driven by Big Data[J]. *Applied Sciences,*2024,14(18):8181-8181.

[24] Hwang, J.-J., Xu, R., Lin, H., Hung, W.-C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., Zhou, Y., Guo, J., Anguelov, D., & Tan, M. (2024). EMMA: End-to-End Multimodal Model for Autonomous Driving. *arXiv*. https://arxiv.org/abs/2410.23262

[25] Pham, T.-H., Ngo, C., Bui, T.-D., Quang, M. L., Pham, T.-H., & Hy, T.-S. (2025). SilVar-Med: A speech-driven visual language model for explainable abnormality detection in medical imaging. *arXiv.* https://arxiv.org/abs/2504.10642

[26] Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., & Wang, X. (2022). GroupViT: Semantic Segmentation Emerges from Text Supervision. *arXiv preprint arXiv:2202.11094.* https://arxiv.org/abs/2202.11094

[27] Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., & Hoi, S. (2021). Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *arXiv preprint arXiv:2107.07651*. https://arxiv.org/abs/2107.07651

[28] Wei, D., Li, Z., & Liu, P. (2024). Omni-Scene: Omni-Gaussian Representation for Ego-Centric Sparse-View Scene Reconstruction. *arXiv preprint arXiv:2412.06273*. https://arxiv.org/abs/2412.06273

[29] Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., & Zhou, J. (2024). RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *arXiv preprint arXiv:2306.11029*. https://arxiv.org/abs/2306.11029