ISSN:2790-1688 Volume-15-(2025)

Vision-Language Models: A Review of Applications and Future Directions

Chenyu Guo

School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China

Email: dutgcy@gmail.com

Abstract. Multimodal artificial intelligence, especially Vision-Language Models (VLMs), has made significant progress in bridging the "heterogeneity gap" between visual perception and natural language understanding. This review aims to comprehensively sort out the core technologies, applications, and future challenges of vision-language models. The article first deeply explores the three key technologies supporting VLMs: multimodal representation learning aimed at building a shared semantic space, modal alignment using mechanisms such as cross-attention to achieve fine-grained correspondence, and hybrid fusion strategies that realize information synergy through in-depth interaction. This paper further outlines the wide applications of VLMs in various fields such as human-computer interaction, content creation, autonomous driving, and medical health. Finally, the article analyzes the current challenges faced by the models in terms of data dependence, interpretability, and computing costs, and looks forward to the future direction of developing next-generation models that are more efficient, controllable, and scalable.

Keywords: Multimodal artificial intelligence, Vision-Language Models(VLMs), Content creation, Autonomous driving, Medical health.

1. Introduction

The cornerstone of multimodal artificial intelligence lies in bridging the "heterogeneity gap" between different data modalities. Information in the real world exists in various forms, such as text composed of discrete symbol sequences and images made up of continuous pixel matrices, which have significant differences in underlying structures and statistical properties. Converting these heterogeneous data into numerical representations (i.e., embeddings) that machines can uniformly understand and compare is a prerequisite for all subsequent tasks (such as alignment and fusion) to achieve more advanced artificial intelligence.

In this field, Vision-Language Models (VLMs), as the core technology connecting images and natural language, are gradually becoming a key driving force for promoting intelligent applications in various industries. Through the joint modeling of visual and language modalities, VLMs can not only realize cross-modal information understanding and generation but also significantly enhance the naturalness of human-computer interaction and the intelligence level of task execution. These models enable intelligent systems (such as virtual assistants) to evolve from merely "understanding language" to being able to "see images".

The influence of VLMs has permeated many complex scenarios. In the field of healthcare, they assist doctors in image diagnosis description and pathological analysis by jointly analyzing medical images and medical record texts. In autonomous driving and robotics, VLMs help systems perform semantic interpretation and task planning while perceiving the environment. In addition, they support creative work such as text-to-image generation in the field of content creation, and improve information accessibility services by converting visual information into language descriptions, providing convenience for the visually impaired.

This paper aims to review multimodal artificial intelligence technologies centered on vision-language models. The paper will first delve into three key technologies supporting VLMs: multimodal representation learning, modal alignment, and information fusion. Subsequently, this paper will sort out the specific applications of VLMs in fields such as human-computer interaction, content creation, and healthcare. Finally, it will analyze the current challenges faced by VLMs and look forward to their future development directions.

2. Key Techniques

2.1 Multimodal Representation Learning

Multimodal representation learning serves as the cornerstone of multimodal artificial intelligence, with its core objective being to bridge the “heterogeneity gap” between different modalities of data, such as text and images. Data from distinct modalities exhibit significant differences in underlying structure, statistical properties, and representational forms—for instance, text consists of discrete sequences of symbols, while images comprise continuous matrices of pixels. Therefore, transforming this heterogeneous data into a unified numerical representation (i.e., embedding) that machines can understand and compare is a prerequisite for all subsequent tasks (such as alignment and fusion).

Based on the design philosophy of the representation space, multimodal representation learning can be primarily categorized into two major types: joint representation learning and collaborative representation learning.

2.1.1 Joint Representation Learning

Joint representation learning aims to map information from multiple modalities into a shared and unified semantic space. In this shared space, vector representations of semantically related content from different modalities are close to each other in spatial position. This is like a multilingual dictionary, where the same word (semantics) in different languages (modalities) all point to the same definition.

In recent years, methods based on Contrastive Learning have become the mainstream paradigm for achieving joint representation. The core idea is to ”pull positive samples closer and push negative samples farther away”. The model learns an encoder that brings the representations of matching cross-modal data pairs (for example, a picture of a dog and the description ”a dog running on the grass”) closer in the shared space, while pushing the representations of mismatched data pairs (for example, a picture of a dog and the description ”a plane flying across the sky”) farther apart. A landmark work is the CLIP (Contrastive Language-Image Pre-training)[1] model proposed by OpenAI. By pre-training on a massive amount of (400 million) image-text pairs, CLIP has learned an image encoder (such as ViT) and a text encoder (such as Transformer) respectively. Its training objective is to maximize the cosine similarity between the representations of matching image-text pairs while minimizing the similarity between mismatched image-text pairs. This learning method enables CLIP to have strong zero-shot generalization ability, and it can achieve excellent performance in various image classification tasks without fine-tuning for specific tasks. Following CLIP, Google’s ALIGN[2] model further verified the effectiveness of this idea on larger-scale network data with more noise.

$$L_{constrastive} = -\sum_{(i,t)\in P} \log \frac{exp(sim(f_i, g_t)/\tau}{\sum_{j=1}^{N} exp(sim(f_i, g_t)/\tau} \quad (1)$$

The above formula is a simplified example of the contrastive learning loss function, where fi is the feature of image i, g_t is the feature of text τ, P is the set of positive sample pairs, $sim(\cdot)$ calculates the similarity, and τ is the temperature coefficient.

2.1.2 Coordinated Representation Learning

Unlike joint representation, collaborative representation learning learns separate representation spaces for each modality, but requires these spaces to maintain a certain consistency or correlation in structure. It does not force all modalities to be squeezed into the same space; instead, it seeks a ”coordinated” relationship, similar to two maps of different scales that can be aligned with each other.

A classic method is Canonical Correlation Analysis (CCA) and its deep learning extensions. Deep Canonical Correlation Analysis (DCCA)[3] uses deep neural networks to perform nonlinear transformations on data from each modality and then maximizes the correlation between the two sets of representations after transformation. The goal of this method is to find a set of projections such that the correlation between the projected modal representations is the strongest, thereby achieving

alignment between modalities. The advantage of collaborative representation is that it can better preserve the unique features of each modality, making it more suitable for tasks that do not require direct cross-modal comparison but need to utilize the correlation between modalities.

2.1.3 Other expressions of learning methods

In addition to the two mainstream methods mentioned above, generative models are also used to learn multimodal representations. For example, Multimodal VAEs[4] attempt to learn a unified latent distribution that can generate (or reconstruct) data from multiple modalities simultaneously. Through this process, the model is forced to compress the core semantic information of all modalities into a shared latent variable z, which itself becomes a high-quality multimodal joint representation.

In summary, multimodal representation learning is the first step towards achieving cross-modal understanding. By constructing a shared or coordinated semantic space, it lays a solid foundation for subsequent modal alignment and information fusion. Currently, joint representation learning methods represented by CLIP have gained the most extensive applications in both academia and industry due to their strong generalization ability and easy scalability.

2.2 Modality Alignment

After projecting data from different modalities into a comparable vector space through representation learning, the next key task is modal alignment. The goal of modal alignment is to identify and establish fine-grained correspondences between data elements of different modalities. If representation learning solves the ”what” problem (for example, both this image and this text are related to ”dogs”), then modal alignment aims to solve the ”where” and ”when” problems (for example, which pixel region in the image corresponds to the phrase ”brown puppy” in the text; which time segment in the video corresponds to the ”close the door” instruction in the speech).

2.2.1 Implicit vs. Explicit Alignment

Modality alignment can be divided into implicit alignment and explicit alignment based on its implementation methods.

Implicit Alignment: In this paradigm, the model is not directly trained to find corresponding relationships. Instead, it ”incidentally” learns alignment as a by-product while completing a certain upper-level task (such as cross-modal retrieval or visual question answering). For example, in the CLIP model, although its training objective is to match the entire image with the entire text description, its internal attention layer must implicitly focus on the connections between text keywords and key image regions to make correct judgments. This kind of alignment is unsupervised but usually not precise enough.

Explicit Alignment: This method requires datasets with fine-grained annotations for supervised learning, and the model is explicitly trained to output the corresponding relationships between modalities. A typical task is Visual Grounding, which is to locate the corresponding bounding box in the image based on a given text description (such as ”the girl wearing a red hat on the left side of the picture”). Such tasks directly optimize the accuracy of alignment, resulting in more precise outcomes, but at the cost of requiring expensive manually annotated data.

2.2.2 Alignment based on the attention mechanism

Modern multimodal models mainly rely on the Attention Mechanism to achieve alignment between modalities, especially the Cross-Attention mechanism based on the Transformer architecture [5]. The core idea of cross-attention is to use the representation of one modality as the ”Query” and the representation of another modality as the ”Key” and ”Value”. The model calculates the similarity between the Query and all Keys (i.e., attention scores) to determine how much information should be aggregated from the corresponding Values. This attention score matrix itself intuitively reflects the alignment strength between elements of the two modalities.

For example, in image-text alignment, the embedding of each word in the text can be used as the Query, and the image is divided into multiple patches, with the embedding of each patch serving as

the Keys. In this way, the model can calculate the correlation between each word and all patches in the image, thereby achieving the effect of ”anchoring” text concepts to spatial positions in the image. Many advanced vision-language pre-trained models, such as LXMERT[6] and ViLBERT[7], adopt such co-attention or cross-attention Transformer layers to simultaneously learn intra-modal and inter-modal contextual relationships and alignment.

2.2.3 Other Alignment Techniques

In addition to the attention mechanism, researchers are also exploring other alignment methods. Sequence-based alignment: In processing temporal modalities such as video, audio, and text, classic sequence alignment algorithms like Dynamic Time Warping (DTW) and their variants still find applications. For example, in instructional videos with narration (such as cooking videos), it is necessary to align the steps in the recipe text with the operation segments in the video stream. Transformer models can also effectively solve such problems by learning long-distance temporal dependencies[8]. Optimal Transport (OT): Optimal transport theory provides a more mathematically principled perspective for modal alignment. It treats the element sets of two modalities as two probability distributions and models the alignment problem as finding a ”transport plan” to move the ”mass” of one distribution to another with minimal ”cost”. This method can better handle complex correspondences (such as one-to-many or many-to-one) and has shown potential in tasks like visual localization [9].

In conclusion, modal alignment is a bridge connecting different information sources, enabling models to perform cross-modal fine-grained reasoning. Techniques centered on cross-attention are currently the mainstream and most effective methods for achieving alignment. Successful alignment is an indispensable part of realizing high-quality multimodal information fusion and content generation.

2.3 Information Fusion

Information fusion is the core of multimodal learning. After representing and aligning data from different modalities, the model needs to organically combine this information to achieve the synergy of ”1+1 $>$ 2” . A successful fusion strategy can comprehensively utilize the complementary information from various modalities while suppressing noise and redundancy, thereby making more accurate and robust judgments than any single modality. For example, in video sentiment analysis, fusing facial expressions (visual), speaking tone (audio), and dialogue content (text) can yield more reliable results than analyzing any single modality alone. According to the stage at which fusion occurs, traditional multimodal information fusion strategies can be divided into early fusion, late fusion, and hybrid fusion [10].

2.3.1 Early Fusion

Early fusion, also known as Feature-level Fusion, combines features from different modalities at the initial stage of model processing. The most straightforward method is to concatenate the feature vectors extracted from each modality to form a longer, unified feature vector, which is then fed into subsequent machine learning models (such as support vector machines or multi-layer perceptrons) for processing. Advantages: It is simple to implement and can capture the underlying correlations between modalities at a very early stage. Disadvantages: Poor flexibility: It requires all modal data to be synchronous and complete, making it difficult to handle cases of missing modalities. Heterogeneity issue: Crudely concatenating feature vectors of different natures (such as dimensions, sparsity, and data distribution) may not handle the heterogeneity between modalities well. Curse of dimensionality: The dimension of the concatenated features may be too high, increasing the difficulty of model training.

2.3.2 Late Fusion

Late fusion, also known as Decision-level Fusion, adopts a strategy opposite to early fusion. It first trains a separate model for each modality to obtain respective decision results (such as classification

probabilities or prediction scores). Then, in the final stage, these independent decisions are integrated through certain rules (such as voting, averaging, weighted summation, or training a meta-learner) to draw the final conclusion. Advantages: High flexibility: Each modality uses an independent model, and the most suitable architecture for that modality can be selected. It has good robustness to the problem of modality missing. Simple implementation: The model training processes are independent of each other, making it easy to implement and expand. Disadvantages: It completely ignores the early interaction of modalities at the feature level. The fusion occurs too late, resulting in the model being unable to learn complex and deep correlation relationships between features of different modalities.

2.3.3 Hybrid Fusion

Hybrid fusion (also known as intermediate fusion) combines the advantages of early fusion and late fusion, and is currently the most mainstream and effective method in the era of deep learning. It does not perform fusion at a single level, but rather conducts continuous and in-depth interaction and fusion at multiple levels of a deep network model.

The rise of this strategy is closely linked to the popularity of the Transformer architecture. Transformer-based multimodal models, such as ViLBERT[7] and LXMERT[6], typically adopt a dual-stream architecture. Text and images are processed in their respective Transformer encoder streams, but there are multiple co-attentional or cross-attentional layers between these two streams. In these layers, the representations of one modality periodically interact with those of the other modality, mutually updating and enriching each other. This process runs through the entire model, achieving in-depth information fusion from shallow to deep layers and from local to global levels.

More advanced recent models, such as DeepMind's Flamingo[11], demonstrate a more sophisticated fusion approach. It "injects" the output of a pre-trained powerful vision encoder into multiple layers of a pre-trained, "frozen" large language model (LLM) through cross-attention layers. This method effectively integrates visual information into the language processing flow, achieving strong few-shot cross-modal understanding and generation capabilities. Advantages: It can learn complex, non-linear deep interaction relationships between modalities, making it the most performant fusion strategy currently available. Disadvantages: The model structure is complex and has very high requirements for computing resources.

To summarize, information fusion technology has evolved from simple feature concatenation and decision voting to the current deep interactive fusion centered on Transformer. It is this advanced hybrid fusion strategy that enables modern multimodal models to truly integrate multi-source information and achieve in-depth understanding of complex scenarios.

3. Applications in Complex Scenarios

3.1 Human-Computer Interaction & Virtual Assistants

Vision-language models enable virtual assistants to not only "understand words" but also "see pictures". For example, interaction systems based on VLM can achieve natural interaction through image understanding and dialogue generation. For instance, when a user uploads a photo, they can obtain scene descriptions, object recognition, and decision-making suggestions[12, 13]. Such systems are widely used in intelligent customer service, Visual Question Answering (VQA), and AR/VR interaction scenarios, significantly improving interaction efficiency and user experience.

3.2 Content Creation & Media Entertainment

In the field of content production, VLMs support generating text descriptions from images (captioning), generating visual materials from text (text-to-image/video generation), and even conducting creative design based on multimodal prompts[14, 15]. This provides the film, television, advertising, and game industries with automated script writing, material generation, and plot planning capabilities, lowering

the threshold for creation, and promotes personalized content distribution.

3.3 Autonomous Driving & Robotics

In the fields of autonomous driving and robotics, VLMs can assist systems in performing semantic interpretation and task planning while perceiving the environment. For example, through multimodal prompts, robots can complete a ”vision-language-action” closed loop in complex scenarios: identifying scene objects, understanding natural language instructions, and generating executable operation steps[16, 17]. This is particularly crucial for road condition analysis and automatic labeling of driverless vehicles, as well as for service robots to perform complex operational tasks.

3.4 Healthcare

VLMs show great potential in the automatic generation of medical images and clinical reports. By jointly modeling medical images and medical record texts, VLMs can assist doctors in image diagnosis description, lesion localization, and multimodal pathological analysis[18]. In addition, they can provide patients with image-based health Q&A services, improving diagnosis and treatment efficiency and reducing the burden on doctors.

3.5 Education & Accessibility

In education and accessibility applications, VLMs can convert visual information into natural language descriptions, helping visually impaired individuals understand their surrounding environment. At the same time, they can also assist intelligent teaching platforms in realizing ”text generation from images” and ”image-text mutual translation,” providing students with cross-modal learning resources[11]. These technologies not only enhance the intelligence and personalization of educational resources but also promote information accessibility and inclusivity.

4. Challenges and Future Directions

Although Visual Language Models (VLMs) demonstrate strong cross-modal understanding and generation capabilities in multimodal artificial intelligence, their large-scale deployment still faces numerous challenges. Firstly, the bottleneck of data and annotation remains unsolved: existing VLMs rely on massive amounts of high-quality image-text pairs, but data in real-world professional fields (such as medicine and industrial inspection) is scarce and the cost of annotation is high. Secondly, the interpretability and robustness of the models are insufficient: VLMs often struggle to clearly explain their cross-modal reasoning processes and have limited adaptability to noisy inputs or cross-domain migration[13]. Thirdly, the refinement and dynamics of multimodal alignment remain a bottleneck: in open environments, the correspondence between image and language semantics may be highly dynamic, while existing models still mostly rely on static alignment mechanisms[11]. In addition, computational and energy costs restrict the practical implementation of VLMs, especially for real-time applications on resource-constrained devices[17].

Future research should focus on the following directions: lightweight and efficient reasoning: achieving low-latency, multi-scenario deployed VLMs through parameter compression[19], knowledge distillation, and edge computing optimization[20]. Cross-modal self-supervised and few-shot learning: developing training strategies that do not require expensive annotations to improve adaptability to domain data[21]. The development of decision paths and adjustable output content is crucial for enhancing trust and security. Integration of multimodal knowledge and external tools: combining VLMs with knowledge graphs, search engines, or symbolic reasoning to enhance the model’s knowledge coverage and ability to handle complex tasks.

In summary, future visual language models will not only pursue stronger cross-modal understanding and generation capabilities but also need to balance efficiency, security, and scalability, thereby truly supporting the large-scale implementation of multimodal artificial intelligence in complex applications such as healthcare, education, and autonomous driving.

5. Conclusion

This paper presents a systematic review of the field of multimodal artificial intelligence centered on Vision-Language Models (VLMs). We explore a series of key technologies ranging from multimodal representation learning, modality alignment to information fusion, which collectively form the pillars of current VLMs. Among them, Transformer-based architectures play a crucial role in enabling in-depth interaction and fusion between modalities, greatly advancing the development of this field.

By examining the applications of VLMs in various scenarios such as human-computer interaction, healthcare, and autonomous driving, it is evident that they have become a core force driving the intelligent transformation of various industries. However, despite the strong cross-modal understanding and generation capabilities demonstrated by VLMs, their large-scale implementation still faces many challenges, mainly including reliance on high-quality annotated data, insufficient model interpretability and robustness, and high computational resource requirements.

Looking to the future, research on vision-language models will not only pursue stronger performance but also need to balance efficiency, safety, and scalability. Developing lightweight models, enhancing model interpretability and controllable generation capabilities, and integrating external knowledge and tools will be the focus of future research. It is foreseeable that as these challenges are gradually overcome, VLMs will truly support the large-scale deployment of multimodal artificial intelligence in a wider range of complex applications, profoundly changing the way humans interact with the information world.

References

[1] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: International Conference on Machine Learning. 2021. url: https://api.semanticscholar. org/CorpusID:231591445.

[2] Chao Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. 2021. arXiv: 2102.05918 [cs.CV]. url: https://arxiv.org/abs/2102.05918.

[3] Galen Andrew et al. “Deep canonical correlation analysis”. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. ICML’13. Atlanta, GA, USA: JMLR.org, 2013, III–1247–III–1255.

[4] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint Multimodal Learning with Deep Generative Models. 2016. arXiv: 1611.01891 [stat.ML]. url: https://arxiv.org/abs/1611. 01891.

[5] Ashish Vaswani et al. “Attention is all you need”. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. isbn: 9781510860964.

[6] Hao Tan and Mohit Bansal. “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). Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5100–5111. DOI: 10. 18653/v1/D19-1514. url: https ://aclanthology.org/D19-1514/.

[7] Jiasen Lu et al. “ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks”. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2019.

[8] Antoine Miech et al. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. 2019. arXiv: 1906.03327 [cs.CV]. url: https://arxiv.org/abs/1906.03327.

[9] Gabriel Peyr´e, Marco Cuturi, and Justin Solomon. “Gromov-wasserstein averaging of kernel and distance matrices”. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. ICML’16. New York, NY, USA: JMLR.org, 2016, pp. 2664–2672.

[10] Tadas Baltruˇsaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal Machine Learning: A Survey and Taxonomy”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 41.2 (2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.

[11] Jean-Baptiste Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. 2022. arXiv: 2204.14198 [cs.CV]. url: https://arxiv.org/abs/2204.14198.

[12] Junnan Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. 2022. arXiv: 2201.12086 [cs.CV]. url: https://arxiv.org/abs/2201.12086.

[13] OpenAI Josh Achiam et al. “GPT-4 Technical Report”. In: 2023. url: https://api.semanticscholar. org/CorpusID:257532815.

[14] Aditya Ramesh et al. “Hierarchical Text-Conditional Image Generation with CLIP Latents”. In: ArXiv abs/2204.06125 (2022). url: https://api.semanticscholar. org/CorpusID: 248097655.

[15] Kaiyang Zhou et al. “Learning to Prompt for Vision-Language Models”. In: International Journal of Computer Vision 130.9 (July 2022), pp. 2337–2348. ISSN: 1573-1405. DOI: 10.1007/s11263-022-01653-1. url: http://dx.doi.org/10.1007/s11263-022-01653-1.

[16] Peter Anderson et al. “Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments”. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, pp. 3674–3683. DOI: 10.1109/CVPR.2018.00387.

[17] Danny Driess et al. “PaLM-E: an embodied multimodal language model”. In: Proceedings of the 40th International Conference on Machine Learning. ICML’23. Honolulu, Hawaii, USA: JMLR.org, 2023.

[18] Benedikt Boecking et al. “Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing”. In: Computer Vision–ECCV 2022. Springer Nature Switzerland, 2022, pp. 1–21. ISBN: 9783031200595. DOI: 10. 1007/978-3-031-20059-5_1. url: http ://dx.doi.org/10.1007/978-3-031-20059-5_1.

[19] Yi Liu et al. MagicVL-2B: Empowering Vision-Language Models on Mobile Devices with Lightweight Visual Encoders via Curriculum Learning. 2025. arXiv: 2508. 01540 [cs.CV]. url: https ://arxiv.org/abs/2508.01540.

[20] Tongtian Yue et al. LaVi: Efficient Large Vision-Language Models via Internal Feature Modulation. 2025. arXiv: 2506.16691 [cs.CV]. url: https://arxiv.org/abs/2506.16691.

[21] Yuanze Hu et al. TinyAlign: Boosting Lightweight Vision-Language Models by Mitigating Modal Alignment Bottlenecks. 2025. arXiv: 2505. 12884 [cs.LG]. url: https: //arxiv. org/abs/2505.12884.