Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 1, 2025

Implementation Paths of Generative AI in Multimodal Learning

Bin Tan

School of Civil Engineering, Xi'an University of Architecture & Technology, Xi'an 710055, China

Abstract: When AI is currently applied to complex scenarios, it only processes single information such as text and images, which cannot meet people's needs to "perceive things as they do." Instead, multimodal learning, which can integrate different information such as text, images, audio, and video, has become a key direction for the development of generative AI. However, the current multimodal applications of generative AI are still stuck in several hurdles: First, the different modalities are quite different. For example, text is segmented and images are pixelated, which can easily cause problems when put together; second, the generated content does not match the intended meaning of the multimodal representation; and third, the quality of cross-modal generation is also unstable, sometimes good and sometimes bad. The stakeholders outlined its implementation approach across five key aspects: data foundation, feature processing, model architecture, training strategy, and quality assessment. First, pre-process and standardize multimodal data to establish a solid data foundation; then, cross-modal feature alignment is used to address modality differences; then, the generative model architecture is adapted to support cross-modal generation; then, multimodal pre-training and incremental learning are used to adapt the model to a wider range of scenarios; and finally, a scientific quality assessment system is employed to optimize the generation results. This research aims to provide usable technical logic for the implementation of generative AI in multimodal scenarios such as intelligent interaction, content creation, and medical diagnosis, and to promote multimodal generation from "being able to generate" to "generating well"

Keywords: Generative AI, multimodal learning, modal alignment, cross-modal generation, pre-trained model, generation quality assessment.

1. Introduction

Human cognition of the world is inherently a multimodal fusion process—visual perception of images, auditory perception of sounds, and language-mediated text, which are then integrated into the brain to form a complete cognition. This has driven AI's transition from unimodality to multimodality: Initially, GPT-2 processed only text, and GANs generated only images. By 2022, GPT-4V achieved cross-modal understanding of text and images, and by 2023, DALL-E 3 was capable of text-guided, detailed image generation. The multimodal capabilities of generative AI have become a core indicator of intelligent development. The multimodal applications of generative AI have long permeated numerous fields. Decisions made by autonomous driving systems require a comprehensive combination of camera images, radar data, and voice commands; medical diagnoses rely on the collaborative analysis of CT scan images and medical records; and content creation can be driven by text to generate images, audio, and short videos. However, these multimodal applications still face significant technical challenges. First, modal heterogeneity is prominent. Text is segmented into word units, and images are composed of pixels, making direct feature fusion difficult. Second, semantic bias is easily present in the generated data, such as the text "red flowers" corresponding to a blue image, or the audio and video rhythms being misaligned. Third, the model lacks scalability. After adapting the text-image combination, adding audio can lead to a significant performance drop. This study did not use fictitious data. By combing through public literature, analyzing open-source models such as CLIP, and analyzing industry practices, it constructed a logical chain for technology implementation from five dimensions: data, features, architecture, training, and evaluation, providing a reference for related research and engineering practice.

2. Multimodal Data Preprocessing and Modality Standardization: Building a Reliable Data Foundation

To do multimodal learning well, you must first have high-quality data. The success of generative AI depends entirely on whether the input data is complete and consistent. If the data is noisy, such as blurry images, incorrectly labeled text, or missing modalities or messy formats, it will be difficult to advance feature fusion and content generation. Therefore, multimodal data must be preprocessed and standardized. This is the first step. The core is three things: data selection, format conversion, data cleaning and alignment. When selecting data, you should choose data with complete modalities and accurate annotations, whether it is public or scenario-specific. Microsoft COCO has 330,000 images with corresponding text annotations, and even the object category and location are clearly marked. It is the foundation for text-image generation. Flickr30K has 300,000 images and 1.5 million texts, with 5 texts for each image, which can help the model learn multiple semantics. LAION-5B is an open source large data set with 5.5 billion sets of image-text pairs, which can support large-scale pre-training [1]. In practice, you should choose according to the task, such as Medical ImageNet for medical treatment and LibriSpeech for audio-text generation. Format conversion involves aligning different modalities into a unified format that the model can recognize: text is first segmented, converted into word vectors, and then unified in length; images are aligned in length and width, and pixel valuesare standardized; audio is converted to mel-spectrograms, and features are extracted to generate fixed-dimensional vectors. For example, when CLIP processes text

and images, the text uses a Transformer encoder and the images use a ResNet or ViT encoder, ultimately converting both into 768-dimensional vectors, thus achieving a unified format. Data cleaning and alignment are the final steps to ensure quality. Data cleaning involves removing poor samples: blurry or heavily occluded images, mislabeled or inappropriate text, and noisy or short audio. Generative AI multimodal applications have already entered numerous fields: autonomous driving relies on images, radar, and voice for decision-making, healthcare relies on CT scans and medical records, and creative writing generates images and audio from text. However, there are technical obstacles here, as the analysis literature and other materials involve five fields.

3. Designing a Cross-Modal Feature Alignment Mechanism: Overcoming the Challenge of Modal Heterogeneity

In generative AI, the most difficult aspect of multimodal learning is modal heterogeneity—text is segmented, images are stacked pixel by pixel, and audio is linked to time. These features differ significantly, so simply combining them can easily lead to errors, resulting in mismatched generated content and meaning. For example, if text and image vectors are simply concatenated, the model may not understand which pixel region in the image corresponds to "red," resulting in incorrect semantics. Therefore, cross-modal feature alignment is necessary to bring features from different modalities into the same semantic space. Currently, three main approaches are available. Shallow feature alignment focuses on placing features in the same space and adjusting them linearly or nonlinearly to make low-dimensional features comparable. The classic OpenAI CLIP approach relies on contrastive learning: text and images are each passed through an encoder for translation, and the model is trained using a contrastive loss function to keep matching image-text pairs close together and mismatching pairs further apart. After training, shallow alignment can be achieved to find similar images for text and assign captions to images. This approach is low-cost and easy to implement, and is commonly used for text-to-image search and image-to-caption generation. Deep semantic alignment places more emphasis on the semantic connection between modalities. It is not just about the numerical similarity of features, but also about making the model understand that different modalities are talking about the same thing. The mainstream relies on cross-modal attention mechanisms to build interactions, with Meta's BLIP and Google's FLAVA being representatives. Take BLIP as an example. An attention layer is added between the text and image encoders: image features are used as "keys" based on pixel blocks, and text features are used as "query terms" based on words. The weights are used to make the words correspond to the image areas. For example, in the text "a red apple is placed on a white plate", "red" focuses on the apple area and "white" focuses on the plate area. This can lay the foundation for text-guided image generation. BLIP-2, combined with a large model, can also write text based on images and answer questions. If you want to add new modalities such as audio to the text-image model, there is a modality adapter method: add a small adapter to the existing bimodal model to specifically handle the feature conversion and semantic correspondence of the new modality, without having to rebuild the model [2]. For example, by adding an audio adapter to CLIP and transferring the audio features into a unified semantic space, a small amount of data fine-tuning can achieve trimodal alignment, which reduces costs while ensuring performance. It is now used in multimodal content search and intelligent voice interaction.

4. Multimodal Adaptation of Generative Model Architectures: Enabling Cross-Modal Generation Capabilities

The key to generative AI is to produce content that matches the meaning, but the old model architecture is too simple - GPT only processes text, and GAN only generates images, and cannot directly handle multimodal tasks. For example, GPT's autoregressive structure can only process a string of data and cannot resolve the spatial features of the image; although GAN can generate images, it is difficult to incorporate text semantics into the guidance. Therefore, the model architecture must be adapted to enable it to have a one-stop capability from feature fusion to cross-modal generation. Currently, there are three mainstream adaptation methods. The multimodal extension of the Transformer architecture is the most widely used. The key is to rely on the "multimodal encoder-decoder" to fuse the generated features together. A typical example is GPT-4V. Based on the original text encoder and decoder, a ViT-L/14 visual encoder is added. The image is cut into 16×16 pixel blocks, converted into visual features, and then fed into the decoder through an adapter and text features. This architecture can uniformly process sequence data and can also continue to use the advantages of GPT-4 to write text, allowing images to guide the writing of text and text to guide the understanding of images [3]. Meta's LLaVA is similar. It uses an adapter to connect the CLIP visual encoder and the LLaMA text decoder. It is open source and can be searched. It can write creative text based on images and solve image-related math problems. The diffusion model mainly performs text-guided image and video generation. The key is to integrate text semantics into the "conditional diffusion process". The diffusion model originally generates clear images from noise. The focus of the adaptation is to make the diffusion process follow the text. For example, Stable Diffusion adds the CLIP text encoder and cross-attention layer to the traditional diffusion model. It first converts the text into a semantic vector. In each diffusion step, the cross-attention layer allows the image features and text semantics to interact. For example, if the text says "Impressionistic seaside town under the sunset", it will guide the generation of an orange-red, wavy picture. This architecture has high generation quality and is easy to control. It can be used for text-to-image, style transfer, and short video generation. The open source community also has plug-ins such as ControlNet to help with precise control. The autoregressive model is mainly used for sequence generation such as text-to-audio and text-to-video. The key is to convert non-sequential modalities into token strings through "multimodal tokenization" [4]. For example, Google's AudioLM uses an audio tokenizer to convert audio into discrete tokens, which are then decoded into audio according to the text generation logic. This approach is logically sound and suitable for long content such as audiobooks. Meta's Make-A-Video converts video frames into visual tokens, combines them with text token generation, and implements

the "text → frame sequence → video" process. It can now generate 16-frame short videos with content that matches the text very well.

5. Multimodal Pre-training and Incremental Learning Strategies: Improving Model Generalization and Scalability

Generative AI multimodal learning, when trained from scratch, requires not only massive amounts of labeled data but is also prone to overfitting—it can only generate content from the training scenario and fails when encountering new scenarios. For example, a model trained solely on animal images and text would struggle to generate architectural images or understand architectural text. Therefore, multimodal pre-training and incremental learning are essential. First, let the model learn the basics of multimodality from a large amount of general data, and then rely on incremental training to adapt to specific scenarios, improving generalization and the ability to add new modalities. There are actually three core methods. Multi-task pre-training is actually to lay the foundation for the model's general capabilities. The key is to let the model learn multiple multimodal tasks at the same time during pre-training, and to understand the ins and outs of modal fusion. If you look at Google's FLAVA model, you will know that its pre-training tasks are very comprehensive - there are both single-modal understanding such as text, image, and audio classification, as well as cross-modal interactions such as image-text comparison, generation, and audio and video matching. The dataset covers four modalities, and the sample size is as high as hundreds of millions. After training, the model can understand the basic logic of multimodal combination and can be used in specific tasks with a little adjustment. It can also achieve "one model with multiple capabilities", saving the cost of separate training, which is quite cost-effective. GPT-4V and LLaVA are currently using this approach. Modal incremental learning solves the problem of adding new modalities to models. The key is to add a new modality (such as audio) to an existing model (such as a text-image model) with a small amount of data without losing the original capabilities. Traditional full retraining is costly and time-consuming. This method relies on parameter isolation and knowledge distillation for efficient expansion: add an audio encoder and adapter to LLaVA, train only the new parameters, and then use knowledge distillation to transfer the original model capabilities to maintain performance. Open source Audio-LLaVA is like this. Adding the CLAP audio encoder to LLaVA and only training the adapter can achieve trimodal interaction and perform multiple tasks at only 1/5 the cost of full training. Scenario incremental learning mainly adapts to industry scenarios. The key is to fine-tune a general pre-trained model with a small amount of scenario data so that the model can generate content that meets the requirements [5]. For example, a medical multimodal model must be fine-tuned with CT images and diagnostic text on Stable Diffusion to retain general capabilities while learning medical professional semantics. After fine-tuning the domestic medical version of Stable Diffusion with 50,000 sets of medical data, it can generate compliant images according to doctors' descriptions without generating non-medical content. This strategy can adapt to the industry at low cost without building models from scratch. It is currently used in the fields of medical care, education, and industrial design.

6. Multimodal Generation Quality Assessment and Optimization: Ensuring the Reliability of Generated Content

Generative AI in multimodal learning must not only be “capable of generation” but also “reliable”. If the generated content does not match the semantics, such as the text says “safe driving” but the accompanying picture is a car accident scene; or the quality is not good, such as the picture is blurry and the audio is full of noise, it is completely unusable. A scientific multimodal generation quality evaluation system must be established and the model must be adjusted according to the evaluation results. This is the key to its use in multimodal scenarios. The core is three steps: designing evaluation indicators, implementing optimization methods, and building feedback loops. Evaluation indicators must take into account both objective and subjective factors, and do not only focus on one side and cause deviations. Objectively, use numbers to measure: for text to image, the higher the CLIP score, the more accurate the semantics, and the lower the FID, the clearer the image; for text to audio, use PESQ and STOI to evaluate clarity and naturalness; for multimodal text generation, use BLEU and ROUGE to check whether it is smooth and complete [6]. Subjectively, rely on user feedback, such as asking doctors to score image-to-text reports in medical scenarios, and asking users to evaluate whether content creation is correct and good-looking. The industry often uses "objective + subjective" methods. For example, when OpenAI evaluated DALL-E 3, it not only used indicators but also collected tens of thousands of user feedback. Targeted optimization strategies should be implemented based on evaluation results. For low CLIP scores, cross-modal attention can be strengthened, such as by adding text semantic weights to the Stable Diffusion cross-attention layer to focus on key information. For high FID, the number of sampling steps in the diffusion model can be increased, from 20 to 50 to reduce ambiguity. For low BLEU scores, semantic coherence constraints can be added to the decoder. The optimization of DALL-E 3 is a typical example. The previous generation DALL-E 2 had problems with poor text detail restoration, such as generating 3 cats out of "5 cats". OpenAI added a "text detail matching rate" indicator, which can accurately restore details after optimization. Building an "evaluation-feedback-optimization" closed loop is the key to iteration. The mainstream approach in the industry now is to incorporate user feedback. Most multimodal applications often implement a binary rating system of "satisfied/unsatisfied" or a quantitative rating system of "semantic match" (1-5 points). They first collect user feedback, filter out unsatisfactory samples, and then fine-tune the training set after annotation. Google's Gemini model, for example, employs this strategy: it specifically collects user text, image, and audio feedback, and adjusts the training set monthly. This effectively addresses scenario-specific issues, adapts to dynamic demand changes, and ensures development progress.

7. Conclusion

Generative AI multimodal learning is supported by a cohesive technical framework centered around "data-feature-

architecture-training-evaluation." Multimodal data preprocessing and standardization provide a solid foundation, addressing issues like low data quality and inconsistent formats. Cross-modal feature alignment mitigates modal heterogeneity and establishes semantic associations between different modalities. Optimizing the generative model architecture improves generation performance and supports cross-modal content generation. Multimodal pretraining combined with incremental learning enhances model generalization and scalability, reducing industry implementation costs. Multimodal evaluation and optimization ensure content reliability, advancing models from "usable" to "useful." These five approaches are not isolated but rather synergistically integrated. For example, high-quality data preprocessing facilitates feature alignment, while evaluation results can in turn optimize data selection and model design, forming a closed-loop technology. Significant progress has been made in the implementation of generative AI multimodal learning: GPT-4V enables deep interaction between text and images, LLaVA lowers the barrier to adoption through open source, and Stable Diffusion promotes widespread text-to-image conversion in the creative sector. However, challenges remain: modality imbalance: graphics and text perform relatively well, while audio, video, and text-3D combinations underperform; weak small-sample generation capabilities, resulting in poor generalization in scenarios like healthcare and industry where labeled data is scarce; and a misalignment of deep semantics, with some models achieving only superficial matches. Future research could focus on three key areas: integrating neural symbolic systems with multimodal generation to enhance deep semantic understanding; optimizing multimodal small-sample learning through meta-learning and transfer learning to reduce data dependency; and establishing a safety and ethical framework to prevent the generation of harmful content. Only through the simultaneous implementation of technological breakthroughs and regulatory improvements can generative AI better serve societal needs and drive the evolution of intelligent applications into complex scenarios that align with human cognition.

References

[1] Suzuki M, Matsuo Y. A Review of Multimodal Deep Generative Models [J]. Advanced Robotics, 2022, 36(5-6): 261-278.

[2] Radford A, Kim JW, Hallacy C et al. Learning Transferable Visual Models from Natural Language Supervision [C] // International Conference on Machine Learning. PmLR, 2021: 8748-8763.

[3] Rombach R, Blattmann A, Lorenz D et al. High-Resolution Image Synthesis Based on Latent Diffusion Models [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10684-10695.

[4] Sun Lifeng, Song Xinhang, Jiang Shuqiang, et al. Preface to the special topic of multimodal collaborative perception and fusion technology [J]. Journal of Software, 2024, 35(05): 2099-2100. DOI: 10.13328/j.cnki.jos.007030.

[5] Chen Gongguan, Liu Hui, Li Hengtai, et al. Research on subgraph matching contrastive learning method for multimodal pre-training [J]. Journal of Computers, 2025, 48(04): 893-909.

[6] Yao J, Hu Y, Yi Y, et al. MMMG: A Comprehensive and Robust Evaluation Suite for Multi-Task Multimodal Generation [J]. arXiv preprint arXiv:2505.17613, 2025.