Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 10
ISET 2025

Generative and Discriminative Models in Multimodal AI: An Analysis of Vision-Language Tasks

Fengjiang He *

Mathematics and Computer Science Department, Drew University, Madison, NJ 07940, USA

* Corresponding Author Email: 18016423036@163.com

Abstract. The transformer architecture has triggered groundbreaking works in multimodal vision and language (V+L). This article offers brief look into the two main modeling paradigms—generative and discriminative—from their roots in natural language processing (NLP) specifically generative pre-trained transformer (GPT) and bidirectional encoder representations from transformers (BERT), respectively. The core ideas of these two paradigms are then examined to show how they have been modified to handle V+L tasks, resulting in different architectural paths and pre-training methods. The paradigms are also surveyed by core dimensions, analyzing the challenges along the path from distributed paradigms to unified models (e.g., model hallucination, limited evaluation capability and scalability). This work aims to provide a well-organized and clear view on how V+L modeling has evolved and possibly evolved into for researchers as well as practitioners.

Keywords: Multimodal AI; Vision-Language Models; Generative Models; Discriminative Models; Transformer; BERT; GPT.

1. Introduction

The recent renaissance of artificial intelligence (AI) can be mostly credited to the transformer architecture, the 'self-attention' mechanism of which revolutionized the sequence modeling task and helped achieve extreme computational parallelism during training [1, 2]. This breakthrough triggered the development of large language models (LLMs) and resulted in a division of natural language processing (NLP) into two leading paradigms: discriminative method, exemplified by Google's bidirectional encoder representations from transformers (BERT), which is beneficial for understanding tasks by training deep bidirectional representations [3], and generative method, characterized by OpenAI's GPT, which is good at generating content through autoregressive text generation [4]. This distinction is based on a more fundamental difference in probabilistic modeling: discriminative models first separate the estimation of the conditional probability $P(y|x)$ from the manipulation of features $\varphi(x)$ to produce the output y, while generative models first separate the estimation of a joint probability $P(x, y)$ (underlying the data distribution) from the conditional probability $P(y|x)$ [5].

With the maturing of unimodal models, this fundamental dichotomy naturally entered the multimodal world, as researchers tried to construct models that could relate visual perception to linguistic comprehension [6]. The main argument of this review is that the generative vs. discriminative framework offers a fruitful perspective to dissect the developmental process, architectural choices and capacity limits of vision-language (V+L) models. Studying how these two frameworks were developed and have more recently come together will help clarify the field's progress to date and its most pressing challenges.

This paper maps this evolution in a systematic manner. The discussion begins by laying the theoretical groundwork for the transformer and the two modeling paradigms, namely the standard approximation and contraction. The progress of both discriminative and generative V+L models is then reviewed. Next, a comparative discussion and analysis is offered, followed by an exploration of the trend toward unification, concluding with the main challenges and future directions that will drive the evolution of next-generation multimodal AI.

2. Theoretical Foundations and Core Architectures

2.1. The Transformer Backbone: Architecture Deep Dive

Self-attention mechanism of the transformer that calculates representations by considering all other elements in the sequence, has enabled the model to capture the complex dependencies without the use of recurrence [7]. For every input, three new vectors are generated: a query (Q), a key (K) and a value (V). The output is a convex combination of the value vectors, weighted by the compatibility of the query with all keys. This function, which we call Scaled Dot-Product Attention is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The scaling factor $\sqrt{d_k}$ (which is the square root of the dimension of the key vectors) is important to ensure gradients are stable during training, by avoiding the dot product growing unnormalised in magnitude [8]. This core mechanism is then suggested to be extended to multi-head attention for generating multiple linear projections which are used to project the sentence representation at different dimensions simultaneously [9]. Finally, as the architecture is purely transverse without any recurrence or convolution, positional encodings are included in the input embeddings by summing an appropriate position encoding matrix to the embedding of each token, in order to inject some information about the absolute or relative position of the tokens in the sequence. This architecture, which originally consisted of an encoder and decoder, resulted in the two prevailing models.

2.2. The Discriminative Paradigm: BERT's Bidirectional Encoding

The discriminative approach like BERT applies transformer's encoder layer to learn deep bidirectional representations. Its main pre-training task, masked language modeling (MLM) objective, is to corrupt a random subset of the input tokens and then to predict these masked tokens from their full (left and right) context [10]. The resulting "fill-in-the-blank" task provides an ideal training signal, making it an efficient approach for a wide range of sequence-to-sequence, relevance prediction, and language generation tasks. At a more probabilistic level of description, these models are remarkably good at estimating the conditional probability P(y | x).

2.3. The Generative Paradigm: GPT's Autoregressive Decoding

On the other hand, the generative paradigm (e.g., the GPT series) is based on the transformer's decoder architecture. It is trained with a simpler next-token prediction objective into which the model is conditioned to generate the next token of a sequence, conditioning here only on the tokens it has already generated [11]. This is enforced in the self-attention mechanism with a causal mask that does not allow any position to attend to subsequent positions. This autoregressive formulation naturally enables the model to capture the joint probability of data P(x, y), allowing the model to create new, coherent sequences of content. This "architecture destiny" imposed by the specific pre-training objectives has heavily influenced the evolution of V+L models.

3. Discriminative Models in Vision-Language Tasks

3.1. Architectural Evolution and Innovations

Discriminative models in the V+L domain use BERT-like deep contextual model to tasks that need joint reasoning and classification. Their development, as summarized in Table 1, represents a definite trend toward better designs of multimodal fusion and alignment methods. Early influential models like ViLBERT [12] were designed as a two-stream model, with distinct transformer encoders for visual features (usually from an object detector) and text. These parallel streams were not orthogonal: they communicated at selected depths via purpose-designed "co-attentional transformer layers,"

enabling iterative cross-modal grounding. This method was flexible but risked being parameter heavy. Later, LXMERT [13] introduced an even more sophisticated architecture accompanied by three separate encoders - one for language, one for visual objects, and a final cross-modal encoder—to more explicitly decouple intra-modal reasoning and cross-modal fusion, in order to better understand each modality independently before fusion together.

Table 1. A Taxonomy of Discriminative Vision-Language Models.

Model	Year	Core Architecture	Key Pre-training Tasks	Primary Innovation/Contribution
ViLBERT	2019	Two-stream Transformer, Co-attention	MLM, ITM, Masked Region Class.	Pioneered the two-stream pre-training architecture for V+L tasks
LXMERT	2019	Three encoders (Lang, Obj-Rel, Cross)	MLM, ITM, Masked Object Pred.	Designed separate intra-modal and cross-modal encoders
UNITER	2020	Single-stream shared Transformer	Conditional MLM/MRM, ITM, WRA	Adopted a unified architecture; introduced WRA for fine-grained alignment
Oscar	2020	Single-stream, Triplet input	Masked Token Loss, Contrastive	Used object tags as anchors to explicitly ease alignment learning

A notable change of scene to architectural simplification and deeper fusion was the paradigm shift brought by single-stream models such as UNITER [14]. By first passing concatenated visual and textual features through a joint transformer, UNITER allowed visual and textual information to interact earlier and more deeply, and was a smaller number of parameters than a two-stage pipeline. Nevertheless, the key challenge of these models is that matching the textual phrases with the visual regions is intrinsically ambiguous. Oscar [15] proposed a landmark solution by constructing object tags inferred from the image as explicit "anchors". By representing the input as a (word, tag, region) triplet, Oscar made the alignment problem easier: we only need to align words to semantic tags, which are already aligned to image regions. It immensely accelerated learning by making the optimization task eliciting stronger semantic signal and led to state-of-the-art performance on several benchmarks.

3.2. Pre-training Strategies for Multimodal Understanding

The strength of these models comes from pre-training on large-scale image-text datasets using a variety of self-supervised pre-training tasks. These tasks are closely related to the BERT's pretraining targets. The original task is MLM, which is fine-tuned to the multimodal setting: the model should predict masked tokens from the surrounding text and the whole visual input. The visual counterpart of contrastive learning is the masked region modeling (MRM) in different flavors: some models try to regress the masked features and others consider it as a classification task whether the masked object category belongs to a pre-defined vocabulary [16]. Another typical task is image-text matching (ITM), which is formulated as a binary classification task, and the model is trained to predict whether a given image-text pair is matched or mismatched [17]. This promotes a consensus of alignment across the scale of the object, and it is further enhanced by hard negative mining to offer the model harder examples. Subsequent stronger models such as UNITER introduced this as word-region alignment (WRA), leveraging optimal transport theory to explicitly facilitate fine-grained mapping between certain words and particular regions of the image (in contrast to global alignment) [14].

3.3. Applications and Performance Analysis

Following pre-training, these discriminative models are fine-tuned on specific downstream tasks where they have established new state-of-the-art performance. They work in the context of visual question answering (VQA), which they cast as a classification problem integrating the input image with question to predict the answer in a selected set of candidates [18]. In visual common-sense reasoning (VCR), a more difficult task that consists of selecting the reasoning to help with the answer, their deep reasoning model is found to be effective, as they can comprehend the implicit relations between visual elements and textual questions [19]. In the case of the image-text retrieval task, the ITM pre-training task naturally align the models to score similarity between an image or text query and an image or text document for retrieval. For example, models such as Oscar achieved state-of-the-art performance on these tasks, with their large version achieving 73.8% accuracy on the VQA v2 test-std set and greatly enhancing retrieval recall for seen instances of common objects in context (COCO) [15].

4. Generative Models in Vision-Language Tasks

4.1. Architectural Evolution and Innovations

Generative models leverage the inventive powers of GPT on V+L tasks, namely, the task of generating new open-ended text from visual input. Their development, which is detailed in Table 2, has been a progression over time to larger scale, with more efficiency in data, with more generalization. There has been early work that directly adapted pre-trained language models (PLMs) more efficiently [20]. It employed specialized attention processes to incorporate visual information while 4A ensuring that the rich linguistic knowledge contained in the PLM was retained, a core ingredient to avoid "catastrophic forgetting" of useful linguistic priors and to achieve strong performance given a limited amount of in-domain data.

Table 2. A Taxonomy of Generative Vision-Language Models.

Model	Year	Architectural Philosophy	Generation Mechanism	Primary Innovation/Contribution
VisualGPT	2021	Adapt Pre-trained Language Model (PLM)	Autoregressive, injects vision via special attention	Data-efficient PLM adaptation, prevents catastrophic forgetting
SimVLM	2022	Minimalist end-to-end Seq2Seq	Prefix Language Modeling (PrefixLM)	Single PrefixLM objective, scaled with massive weak supervision
Flamingo	2022	Bridge frozen unimodal models	Injects vision into frozen LLM via gated attention	Pioneered effective in-context few-shot learning for multimodal tasks
GPT-4V	2023	Large-scale native multimodal model	Autoregressive generation on interleaved sequences	Achieved powerful zero/few-shot generalization and emergent abilities

A significant leap was achieved with the minimalist frameworks such as SimVLM [21], which gave up complex, multi-stage full-pipeline pre-trainings dependent on object detectors. Instead, it adopted a single prefix language modeling (PrefixLM) [22] objective, which is a simpler and more scalable

counterpart that was trained on end-end on huge, web-crawled data with billions of noisy image-text pairs. The notion of few shot generalization was crystallized by models such as Flamingo [23], which proposed a new architecture to connect strong pre-trained frozen models in vision and language. Its key to success was in two ingredients: a perceiver resampler to compress many visual features into a few information-dense tokens, and new gated cross-attention layers which were interleaved inside the frozen LLM in order to supply this visual context without disturbing the LLM's fine-tuned parameters. This supported in-context learning, and the model was able to execute new tasks based on only a few task-lead examples without any parameter updating. This evolution has led to the current era of large multimodal models (LMMs) such as GPT-4V, which have impressive zero- and few-shot "emergent abilities" across a wide spectrum of tasks not seen during training, including complex chart analysis and understanding visual humor [24].

4.2. Pre-training and Generation Strategies

The heart of all these models is autoregressive generation, i.e., generating words one at a time, conditional on the visual input and all previous tokens generated so far. SimVLM's PrefixLM task is a strong instance, where the model processes a prefix (such as image patches) in both directions to build a rich representation, and autoregressively generates the following text. This marries the cognition ability of encoders with the generation capability of decoders. The most crucial strategic change, however, is in-context few-shot learning, which was pioneered by Flamingo [23]. Rather, one large model can be fine-tuned on a new task at inference time with just a small prompt (a few examples), sidestepping task-specific fine-tuning and bringing general AI closer. The efficacy of this approach depends on the learned-from-web-scale-data model being capable of recognizing patterns and abstracting task instructions from the training examples.

4.3. Applications and Performance Analysis

Generative models excel at tasks requiring unconstrained, open-ended outputs. In image captioning, they generate descriptive sentences for images, which are evaluated by metrics such as consensus-based image description evaluation (CIDEr) [25] and semantic propositional image caption evaluation (SPICE) [26]. They have set the state-of-the-art on image captioning tasks such as COCO captions and NoCaps, and have been observed to generate richer and human-like descriptions than discriminative models. In visual storytelling, they generate consistent narratives out of a series of images, in a setting that requires them to reason about causal and temporal relationships, and is naturally suited to their process of sequential generation. Their most powerful usage is in open-ended VQA where they can produce arbitrary free-form answers instead of selecting from a list. This means they can tackle more challenging questions, which might have needed commonsense facts or to be explained with nuance—a capacity demonstrated impressively by GPT-4V, which was able to give detailed paragraph-length responses based on visual evidence.

5. Synthesis and Comparative Analysis

5.1. The Trend Towards Unification

While V+L research was originally divided along separate discriminative and generative lines, a significant movement has been towards combined models. These paradigms seek to deal with understanding and generation as a single architecture and are of the "one model for all tasks" school of thought. This can be done by enforcing a uniform representation across tasks, such as in the case of one for all (OFA) [27], which uses textual descriptions to define the task, or by unifying the pre-training tasks per se. Models like BEiT-3 [28] belong to the latter category and utilize a single masked "data" modeling objective across modalities over a common multiway transformer backbone. This enables the same model to perform equally well as an encoder, a decoder, or may share a joint model as both based on the downstream task, which is a promising step towards truly general-purpose foundation models.

5.2. Multi-Dimensional Comparison

In comparing the two original paradigms, it is evident there is a trade-off, as summarized in Table 3. For tasks in which classifier or score labeling is needed, discriminative models generally perform better at inference and are more controllable in computation. Their outputs are constrained, so that they are less likely to generate meaningless or ungrounded responses. Generative models on the other hand, are extremely unconstrained in their flexibility, and thus can excel in creative tasks, where the output space is large and ill-defined. But they are autoregressive and require iterative inference (as generation is sequential), and are more prone to hallucinations, a fact when the generated text contains elements that are not in the visual input. Unified models are designed to have their cakes and eat them, too, and it remains to be seen how well they can combat as even the efficiency results for autoregressive generation can be driven by individual creative tasks. Which one is to be used depends primarily on the application needing more accuracy vs. the creative part.

Table 3. Multi-Dimensional Comparison of V+L Paradigms.

Dimension	Discriminative Models (BERT-like)	Generative Models (GPT-like)	Unified Models
Core Capability	**Understanding & Distinguishing**: Learns decision boundaries.	**Creating & Generating**: Learns data distribution.	**Both**: Supports understanding and generation tasks.
Typical Tasks	VQA (Classification), VCR, Retrieval, Grounding.	Image Captioning, Storytelling, Open-Ended VQA, T2I Generation.	Handles all the above and more complex, mixed-modality tasks.
Architectural Focus	Encoder-centric, bidirectional context.	Decoder-centric, autoregressive, unidirectional context.	Encoder-Decoder (Seq2Seq) or flexible multiway architectures.
Inference Efficiency	Generally more efficient (one forward pass).	Slower due to sequential, autoregressive token generation.	Efficiency is task-dependent; generation is still autoregressive.

5.3. The Pivotal Role of Pre-training Data

All V+L models are becoming more and more alike one another not only due to architecture but also scale and character of the pre-training data. The community's transition from small, highly curated, human-annotated collections such as COCO and visual genome to large, weakly supervised, web crawled datasets like LAION [29] (billions of examples) has been the single biggest catalyst for the remarkable zero-shot and few-shot generalization ability demonstrated by modern LMM. This large-scale data exposure is the exposure to the long tail of concepts, cultures and knowledge that is infeasible to represent in hand-crafted datasets. But this data at web scale is also a double-edged sword, as it inherently comes with noise (e.g., mis-matched image-text pairs), society biases and erroneous fact. Models trained on this data faithfully replicate these biases, hence data filtering and debiasing are important areas of study.

6. Key Challenges and Future Directions

6.1. The Pervasive Challenge of Hallucination

Despite this accelerated progress, the field is challenged by great difficulties that constitutes the edge of research. The most problematic is model hallucination, in which models produce text that is not supported by the visual input. This problem can be further decomposed into object, attribute, and relation hallucinations, due to noisy data, a strong language prior dominating a weaker visual signal, and imperfect alignment mechanisms [30]. For example, a model correctly recognizing a kitchen but "hallucinating" a microwave because kitchens and microwaves are highly correlated in its training data, even if no microwave is in the image. This can be alleviated in numerous ways, including better data filtering, altered training objectives that favor faithfulness (such as contrastive decoding), and post-hoc statistical validation.

6.2. The Evaluation Conundrum

Another big problem is the evaluation challenge: classic metrics are not good for new models. VQA accuracy, which frequently uses exact string matches, punishes semantically correct but lexically dissimilar answers (e.g., labeling "the man in the jersey" as incorrect if the ground truth is "the basketball player"). Captioning metrics such as bilingual evaluation understudy (BLEU) [31] and CIDEr, which reward n-gram overlap with reference captions, have limited correlation with factual accuracy and can reward fluent nonsense. This has resulted in the exploration of new paradigms, including: judging from powerful LLMs (LLM-as-a-Judge) or cross verifying the generated captions with VQA models to verify the facts per our data. But these new approaches also come with their own potential biases.

6.3. The Bottleneck of Scalability and Efficiency

The huge scale and efficiency challenges associated with training and deploying these large models impose significant obstacles to research and practical use. Training a foundation model can take millions of dollars in compute, putting it out of reach of a few large labs and giving rise to concerns about the environmental cost of such an extravagant use of resources. This has also spurred a strong interest in approaches to make models more efficient, such as model compression (quantization, pruning), knowledge distillation (training a smaller "student" model to mimic a larger "teacher"), and parameter-efficient fine-tuning (PEFT) techniques (e.g., LoRA [32]) that significantly cut down the cost of tailoring models to new tasks by only modifying a small number of model parameters.

6.4. The Frontier of Generalization and Alignment

Finally, despite their promise in generalization to specific tasks, aligning such vast world knowledge with niche domain specific concepts and constraints presents a challenging "last-mile" problem. A model might have a notion of what a "car" is, but might not be privy to the nuances of the specific types of sensors, and the specific ways in which they fail for an autonomous driving application. This requires a lot of manual prompt engineering or additional tuning, underscoring the need for more scalable alignment methods which are capable of effectively injecting domain-specific knowledge without sacrificing the generality of the model.

7. Conclusion

This review has traced the evolution of vision-language modeling, from the generative and discriminative paradigms, back when they were separated at birth by their NLP nature, to their convergence within modern multimodal systems. It was demonstrated how discriminative models such as Oscar can excel at understanding-based tasks, and how in return generative models such as Flamingo push the frontier of few-shot creation. The direction is rather explicit toward broad, large-scale frameworks like BEiT-3, which take full advantage of massive web scale data for zero-shot

generalization that has never been seen before. But due to some basic hindrances, improvements are checked, involving model hallucination, insufficient evaluation metrics, and proactive computational costs that restrict broader deployment and trustworthiness. The next steps for the community involved are, most likely, in the pursuit of foundation models that are more general, factually grounded, and efficient, ultimately breaking through passive perception to active, embodied intelligence.

References

[1] Soydaner D. Attention mechanism in neural networks: where it comes and where it goes. Neural Computing and Applications, 2022, 34(16): 13371-13385.

[2] Luo Q, Zeng W, Chen M, et al. Self-attention and transformers: Driving the evolution of large language models. In 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), 2023: 401-405.

[3] Charoenkwan P, Nantasenamat C, Hasan MM, et al. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. Bioinformatics, 2021, 37(17): 2556-2562.

[4] Bengesi S, El-Sayed H, Sarker MK, et al. Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEe Access, 2024.

[5] Bernardo JM, Bayarri MJ, Berger JO, et al. Generative or discriminative? getting the best of both worlds. Bayesian statistics, 2007, 8(3): 3-24.

[6] Park SM, Kim YG. Visual language integration: A survey and open challenges. Computer Science Review, 2023, 48: 100548.

[7] Ding Y, Jia M, Miao Q, et al. A novel time–frequency transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. Mechanical Systems and Signal Processing, 2022, 168: 108616.

[8] Shen Z, Zhang M, Zhao H, et al. Efficient attention: Attention with linear complexities. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021: 3531-3539.

[9] Wang Z, Yao K, Li X, et al. Multi-resolution multi-head attention in deep speaker embedding. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 6464-6468.

[10] Chung YA, Zhang Y, Han W, et al. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021: 244-250.

[11] Chen L, Wang Z, Ren S, et al. Next token prediction towards multimodal intelligence: A comprehensive survey. arXiv preprint arXiv:2412.18619, 2024.

[12] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 2019, 32.

[13] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.

[14] Chen YC, Li L, Yu L, et al. Uniter: Universal image-text representation learning. In European conference on computer vision, 2020: 104-120.

[15] Li X, Yin X, Li C, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, 2020: 121-137.

[16] Huang Z, Jin X, Lu C, et al. Contrastive masked autoencoders are stronger vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(4): 2506-2517.

[17] Zhang K, Mao Z, Wang Q, et al. Negative-aware attention framework for image-text matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022: 15661-15670.

[18] Shao Z, Yu Z, Wang M, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2023: 14974-14983.

[19] Khan MJ, Breslin JG, Curry E. Common sense knowledge infusion for visual understanding and reasoning: Approaches, challenges, and applications. IEEE Internet Computing, 2022, 26(4): 21-27.

[20] Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys, 2023, 56(2): 1-40.

[21] Wang Z, Yu J, Yu AW, et al. Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904, 2021.

[22] Jin W, Cheng Y, Shen Y, et al. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. arXiv preprint arXiv:2110.08484, 2021.

[23] Alayrac JB, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 2022, 35: 23716-23736.

[24] Yang Z, Li L, Lin K, et al. The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421, 2023.

[25] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 4566-4575.

[26] Anderson P, Fernando B, Johnson M, et al. Spice: Semantic propositional image caption evaluation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, 2016: 382-398.

[27] Wang P, Yang A, Men R, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International conference on machine learning, 2022: 23318-23340.

[28] Wang W, Bao H, Dong L, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 19175-19186.

[29] Schuhmann C, Beaumont R, Vencu R, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems, 2022, 35: 25278-25294.

[30] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[31] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002: 311-318.

[32] Chavan A, Liu Z, Gupta D, et al. One-for-all: Generalized lora for parameter-efficient fine-tuning. arXiv preprint arXiv:2306.07967, 2023.