Vision-Language Integration in Large Language Models: A Survey of Architectures, Training Paradigms, and Applications

Ruize Hou
College of Engineering
Northeastern University
Boston, MA, United States
hou.ruiz@northeastern.edu

Chenyi Xu
College of Engineering
Northeastern University
Boston, MA, United States
xu.chenyi@northeastern.edu

Indrajeet Roy
College of Engineering
Northeastern University
Boston, MA, United States
roy.i@northeastern.edu

***Abstract*—The integration of vision capabilities into Large Language Models (LLMs) represents a significant advancement in artificial intelligence, enabling systems to process and understand both visual and textual information simultaneously. This survey provides a comprehensive examination of Vision-Language Models (VLMs), focusing on their architectural evolution, training methodologies, and applications. The survey systematically analyzes the progression from early two-stream architectures to modern unified frameworks, examining key developments in model design, pretraining strategies, and cross-modal learning techniques. Specific areas covered are architectural choices that enable effective vision-language integration, training paradigms, and emerging evaluation frameworks. The survey concludes by identifying critical challenges and promising research directions in vision-language integration, particularly in areas of architectural efficiency, scalability, and real-world applications.**

***Index Terms*—Vision Language Models, Large Language Models, Visual Feature Extraction, Convolutional Neural Networks, Vision Transformers, Attention Mechanism, Transformer, Latent Space,Multimodal reasoning, Visual Embeddings, Latent Representation**

I. Introduction

In recent years, Large Language Models (LLMs) have demonstrated revolutionary advances in artificial intelligence by integrating visual capabilities. The development of these visual-language models (VLMs) not only pushes the boundaries of multimodal interaction capabilities but also enables computers to understand both visual and textual information, providing brand new solutions for complex tasks such as image description generation, visual question and answer, and cross-modal reasoning. The core breakthrough in this area lies in the utilization of advanced architectural design and training paradigms to achieve a deep fusion of visual and linguistic features [1] [2] [3].

At the architectural level, VLMs have undergone an evolution from early dual-stream to single-stream architectures. The bi-streaming architecture utilizes separate visual and text encoders to extract information separately and subsequently achieves information alignment through a cross-modal attention mechanism [4] [5]. In contrast, the unified streaming model implements continuous visual and text interactions at each layer of encoding, capturing more fine-grained cross-modal semantic relations through a self-attention mechanism [6]. In addition, the patch-based feature extraction approach replaces the traditional object-based detection technique, significantly improving the efficiency and performance of the model when processing high-resolution inputs [**?**].

The diversity of training paradigms has likewise contributed to the rapid progress of VLMs. Contrastive learning lays the foundation for semantic consistency in multimodal representations by maximizing the similarity of matched graphic pairs and the differentiation of unmatched pairs [7]. The introduction of masking objectives enhances the model's cross-modal learning capability. This approach randomly masks a portion of the input and requires the model to reconstruct it. This process enables the model to extract deep semantics efficiently within a self-supervised learning framework [8]. Furthermore, the generative approach further extends the application boundaries of VLMs, enabling them to accomplish tasks ranging from text-generated images to cross-modal reasoning [9].

This review comprehensively analyzes the key developments in architectural design, training paradigms, and their applications of VLMs. Section II explores in detail the key architectural evolutions in key Vision-Language integration components such as Visual Feature Extraction, Feature Space Organization and Vision-Language Integration Architectures and their core innovations. Section III focuses on the theoretical foundations and practical effects of different training paradigms. Section IV summarizes the performance of VLMs in real-scenario applications.

II. Key Architectural Developments

A. Visual Feature Extraction

Visual feature extraction in Vision-Language (VL) integration transforms raw, pixel-level input into semantically rich, high-dimensional embeddings that can be aligned and integrated with textual representations. Using architectures such as Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and multimodal encoders, this process converts low-level visual details into abstract, conceptually meaningful features. As Large Language Models (LLMs) increasingly handle multimodal inputs, the quality and efficiency of visual feature extraction become critical to their ability to perform cross-modal reasoning, visual grounding, and context-aware inference. By providing concept-level understanding of images, these high-quality representations are essential for achieving strong performance, adaptability, and interpretability in Vision-Language Models (VLMs).

1) ***Object-centric Visual Feature Extraction****:* Object-centric, region-based visual feature extraction is a foundational method in Vision-Language (VL) integration [10] [4]. Its key foundation is the utilization of object detection models—such as Faster R-CNN [11] or Mask R-CNN [12] to generate a discrete set of region proposals within an image. Each region corresponds to a localized entity (e.g., an object or a salient scene element) and is represented by a high-dimensional embedding that captures appearance, attributes, and semantic information [13] [5]. These structured, region-level features act as concept-level abstractions, translating raw pixel data into semantically meaningful units closely aligned with linguistic tokens [10] [14]. By incorporating pretrained object detectors, which have already learned to identify and classify visual elements, this approach leverages prior pretraining domain knowledge and aligns visual features more directly with textual concepts, improving fine-grained grounding, contextual reasoning, and cross-modal understanding [4] [5].

Early object detection architectures such as Faster R-CNN [11] for visual feature extraction utilized a convolutional neural network (ResNet-101) [15] for initial image processing to produce dense feature maps capturing high-level visual patterns.A Region Proposal Network (RPN) [11]—a lightweight neural module—analyzes these feature maps to generate a small set of high-quality candidate bounding boxes, effectively highlighting regions that are likely to contain objects or other salient elements.Each proposed region is then extracted and refined through methods such as Region of Interest Pooling [16] or Region of Interest Align [12], resulting in final region-level feature vectors. These vectors serve as semantically rich, spatially grounded embeddings. When integrated with textual inputs, they enable VLMs to perform more sophisticated multimodal reasoning and interpretation.

The Bottom-Up and Top-Down Attention model [10] introduced a key advancement in Vision-Language Models (VLMs) by refining how object-centric visual features are integrated with linguistic context. Earlier VLMs typically relied on static object-level embeddings produced by object detection architectures [10] [15]. In contrast, the Bottom-Up and Top-Down approach adds a dynamic, context-aware attention mechanism that adjusts these visual representations based on input language.In the bottom-up stage, a pretrained object detector (such as Faster R-CNN with a ResNet backbone) processes the image to identify salient regions, each representing an object or a meaningful scene component. These region-level embeddings function like visual tokens, capturing object, attribute, and relational information in a form that can be aligned with linguistic information.During the top-down stage, the learned attention module—an LSTM guided by language inputs like partial captions or user questions—integrates these visual tokens with the given text. By comparing the language input to each region-level embedding, the attention mechanism assigns attention weights that highlight the most relevant parts of the image. This process selectively emphasizes certain visual features, allowing the model to focus on image regions most relevant to the current linguistic query.By dynamically refining visual representations based on textual context, the Bottom-Up and Top-Down Attention model facilitates more contextually grounded, coherent multimodal reasoning. This improvement in the integration of visual and linguistic information enhances the model's performance on tasks such as image captioning and visual question answering.

ViLBERT [4], LXMERT [5], and UNITER [17] represent key advancements in Vision-Language Models (VLMs) that build upon the object-centric foundation. While earlier approaches paired a static set of region-level object embeddings (generated by a Faster R-CNN detector [11]) with textual inputs, the newer models move towards transformer-based architectures [18] for more robust multimodal integration. Instead of relying on attention modules or sequential processing to combine pre-extracted visual embeddings and textual inputs, ViLBERT [4], LXMERT [5], and UNITER [17] adopt transformer architectures designed to jointly model both modalities. Transformers excel at capturing long-range dependencies and complex interactions using self-attention and cross-attention layers. By representing both text tokens and object-level features as visual tokens within this framework, these models establish a shared embedding space where language and vision can interact and influence each other. While the Bottom-Up and Top-Down approach [10] introduced dynamic attention over static region embeddings, transformer-based VLMs apply iterative layers of attention. This iterative process continually refines and contextualizes the joint representation, allowing textual and visual modalities to guide and reshape each other's encodings. As a result, these models produce contextually enriched multimodal embeddings, leading to improved performance in cross-modal tasks such as image captioning, visual question answering, and multimodal reasoning.

*2) **Patch-based Visual Feature Extraction**:* Patch-based (or patch-level) visual feature extraction represents a fundamental shift from object-centric approaches [3]. Instead of detecting and analyzing predefined objects or object-centric regions defined by bounding boxes [11], [12], this method uniformly divides the image into a fixed grid of patches. Each patch serves as a discrete visual token, and a model encodes these patches into embeddings that represent the entire image at a uniform level of granularity [3], [19]. By utilizing a grid of patches rather than preselected objects, Vision-Language Models gain more flexibility [20]. They can dynamically focus attention on any part of the image, rather than being limited to regions identified by an object detector. This patch-based methodology supports a more fine-grained, context-aware exploration of the visual scene [1], allowing models to capture subtle details and relationships that object-centric methods might overlook.

Object-centric (Region-based) visual feature extraction utilizes object detectors—such as R-CNN, Faster R-CNN, or Mask R-CNN operating within a convolutional framework wherein stacked convolutional layers capture hierarchical local spatial patterns, and a region proposal module identifies candidate bounding boxes likely to contain salient objects. Specialized modules classify and refine these proposed regions generating object-level feature embeddings [12] [10]. While this approach is effective for object detection and classification, certain limitations and constraints exist within the methodolgy. The rigid structure of predefined steps (proposal generation, region pooling, classification) resulted in limited adaptability and end-to-end optimization [11]. A constrained pre-training object vocabulary may reduce the model's capacity to generalize beyond known categories, inhibiting recognition of novel or subtle visual information [10] [21]. Convolution-based feature extraction lacks the dynamic flexibility to reweight attention across an image, limiting the model's ability to emphasize different regions dynamically, binding the model to predefined concepts and localized feature maps rather than a more context-aware understanding of visual input [22] [23].

The Vision Transformer (ViT) [3] marks a notable advancement in visual feature extraction by implementing patch-level tokenization in combination with a Transformer [18]. Unlike previous methods that relied on object-centric processes (e.g., bounding boxes, region proposals) or convolutional filters—both imposing strong inductive biases about image structure [11] [24], ViT treats images as sequences of uniformly partitioned patches. This patch-level tokenization treats every portion of the image as an equally valid source of information, eliminating the need for predefined object annotations and manually created semantic boundaries.

[3] In the Vision Transformer (ViT) , uniform patch tokenization is performed by evenly subdividing the input image of dimension $H \times W \times C$ into a grid of non-overlapping patches of a fixed size.The input image is partitioned into $\frac{H}{P} \times \frac{W}{P}$ patches, each of size $P \times P \times C$. These patches are then flattened into 1D vectors and linearly projected into a common embedding dimension, ensuring that each patch is represented as a discrete token.ViT adds learnable positional embeddings to each patch embedding to incorporate positional information as spatial context is lost once the image has been fragmented into tokens. [18] This process transforms the image into a sequence of patch-level embeddings with positional encodings, which is then passed to a Transformer encoder.By applying self-attention over these patch tokens, the ViT model can learn to focus on the most relevant image regions for a given task, [25] eliminating the need for predefined object boundaries, specialized detectors, or convolutional filters.

[25] By relying on patches rather than explicit object proposals or convolutional filters, ViT enables end-to-end training and reduces reliance on external proposals, bounding boxes, or explicitly annotated object categories. [3] The representation is consistent across different input sizes and aspect ratios, simplifying adjustments for new tasks or datasets. As a result, ViT can scale more naturally, allowing the Transformer to learn directly from raw pixel patches without requiring predefined categories or specialized modules.

[3] A key advantage of the patch-based visual feature extraction methodology is its natural alignment with the token-based representations utilized in language models. By representing both images and text as sequences of tokens—patch embeddings for images and word or subword embeddings for text—both modalities share a common, uniform input structure, allowing for a single [18] [20] Transformer to process visual and linguistic data using the same self-attention mechanisms, eliminating the need for different feature extraction methods or specialized modality bridging or alignment mechanisms. [7] The model considers text and images as different modalities within the same tokenized representation space instead of fundamentally dissimilar data types, encouraging the model to learn contextually informed alignments between patches and words assisting in multimodal reasoning and understanding.

*3) **Hierarchical-Patch-based Visual Feature Extraction**:* Hierarchical Patch-Based Visual Feature Extraction extends the patch-level methodology utilized in Vision Transformer (ViT)-based architectures by introducing a structured, multi-scale image representation process across multiple model stages. [26] Rather than relying on a single spatial resolution of uniformly sized patches, this approach progressively refines or aggregates patch embeddings as they advance through successive transformer encoder layers. [26] Each encoder stage operates at a distinct spatial scale or token resolution, constructing a hierarchical latent space where early layers focus on encoding fine-grained local features, while deeper layers integrate these representations into more abstract, semantically rich embeddings. By extracting and refining visual features at multiple scales, the model preserves the modality-

independent advantages of patch tokenization while simultaneously leveraging the multi-scale, context-sensitive properties of convolutional neural networks (CNNs). This hierarchical integration of local and global contexts enhances the model's adaptive visual reasoning, enabling it to capture intricate visual details at lower layers and form higher-level, domain-relevant abstractions at upper layers.

[26] The Swin Transformer represents a significant progression in the patch-based visual feature extraction paradigm introduced by the Vision Transformer (ViT). While ViT processes images as sequences of uniformly sized patches and applies global self-attention at every layer—an approach that becomes computationally inefficient for high-resolution inputs—Swin addresses these limitations by implementing a hierarchical patch representation and localized attention mechanisms. This hierarchical architecture reduces computational complexity, enhances scalability and also facilitates the construction of contextually enriched and semantically structured visual representations. As a result, the Swin Transformer's architecture is well-suited for multimodal reasoning as efficient integration of visual and linguistic information is essential.

The Swin Transformer progressively merges patches across successive stages, unlike the Vision Transformer (ViT), which maintains a fixed patch size and resolution throughout its entire architecture [26] [3]. Initially, the Swin Transformer operates on a fine-grained grid of patches directly extracted from the input image. As the network advances, these patch embeddings are concatenated and linearly projected, effectively increasing the spatial resolution of each patch and simultaneously reducing the token count along each spatial dimension at each merging step [27].This hierarchical downsampling similar to spatial pooling or striding in convolutional neural networks (CNNs) produces a multi-level representation in which early layers encode fine-grained local patterns, while deeper layers capture increasingly global and semantically enriched features [23] [26].By representing images at multiple scales and progressively merging patches, the Swin Transformer produces a more semantically structured visual feature hierarchy, which is effective for multimodal reasoning in vision-language tasks. At lower levels, the model preserves detailed local information that can be directly associated with specific linguistic tokens, facilitating explicit mappings between textual elements and their corresponding image regions, attributes, or objects. At higher levels, the more abstract visual embeddings naturally align with broader semantic concepts such as scene categories, object relationships, and thematic contexts [1]. This multi-scale alignment enables the model to transition from pixel-level details to generalized semantic representations, enhancing contextually grounded and semantically rich multimodal reasoning.

Large-scale multimodal systems, exemplified by PaLM-E, represent a further advancement in hierarchical patch-based feature extraction and vision-language integration. While architectures such as the Swin Transformer establish robust, multi-level visual representations that readily align with linguistic inputs, PaLM-E extends these principles through increased model capacity and data complexity. By integrating hierarchical, patch-based embeddings with large language models (LLMs), PaLM-E effectively processes and infers over multimodal inputs, capturing more intricate relationships between visual features and linguistic context. This approach generates richer and more semantically sophisticated multimodal representations, enhancing grounded language understanding, enabling zero or few-shot inference, and improving cross-modal reasoning capabilities [26] [28] [29].

[29] PaLM-E advances hierarchical patch-based feature extraction by utilizing a pretrained vision encoder—commonly a Vision Transformer (ViT) or a hierarchical variant such as the Swin Transformer—that operates on patch-level embeddings. The encoder segments the input image into a grid of non-overlapping patches and transforms each patch into a visual token via flattening and linear projection. In a standard ViT, global self-attention is applied uniformly across all patches, whereas hierarchical architectures like the Swin Transformer progressively merge patches at deeper layers [3] [26].This hierarchical merging produces a multi-scale representation that transitions from low-level visual features, such as textures and colors, in the early layers to semantically enriched, context-sensitive abstractions, consisting of objects and spatial relationships, at higher layers. This layered image representation enables more flexible and adaptive visual encodings, with lower layers capturing granular pixel-level information and upper layers forming conceptually grounded, context-aware representations.

By projecting hierarchically structured visual embeddings and linguistic tokens into a unified latent space, PaLM-E constructs a joint representation that enables the pretrained PaLM language model to semantically ground and contextualize visual features. Within this shared representation, low and mid-level image attributes (e.g., texture, color, geometric configurations) are mapped onto higher-level semantic and relational categories acquired through large-scale language pretraining [28] [29]. The hierarchical organization of PaLM-E's vision encoder allows the language model to dynamically query and integrate information across multiple abstraction levels. At the lowest layers, fine-grained patch embeddings support detailed perceptual inferences, while intermediate layers merge patches into object-level tokens that serve as discrete, semantically meaningful units for language-guided reasoning. At the highest layers, the model captures broader contextual relationships and spatial arrangements, which can be aligned with conceptual linguistic information to facilitate more advanced, context-aware inferences. This hierarchical encoding and unified token-based multimodal alignment enable PaLM-E to leverage linguistic information for positional visual reasoning, enabling coherent and semantically grounded multimodal reasoning that integrates pixel-level information

with higher-order visual abstractions. [28] [29]

B. Feature Space Organization

Feature Space Organization is the design and structuring of the representational spaces in which different data types such as images, text, or multimodal content are embedded and compared. In Vision-Language Models, feature space organization is critical consideration as it determines how visual and linguistic information are aligned, integrated, and leveraged to achieve sophisticated reasoning capabilities. By establishing a shared embedding space where text and image features can naturally interact, models are better equipped to handle tasks mulitmodal tasks such as image captioning, visual question answering, and scene understanding. This alignment also improves the model's ability to generalize across domains, as a coherent feature space facilitates transfer learning and adaptation to new tasks with minimal retraining.

1) ***Distinct Modal Feature Spaces****:* Distinct Modal Feature Spaces represent a paradigm in which visual and textual modalities are mapped into seperate, unimodal embedding spaces. Under this framework, image and text data streams are processed in isolation: a vision encoder (e.g., a Convolutional Neural Network or Vision Transformer) projects visual inputs into a dedicated latent vector space, while a language model encodes linguistic tokens into a separate, non-overlapping embedding space. By deferring cross-modal integration, this approach preserves modality-specific semantic representations, structural priors, and inductive biases.Vision-Language Models (VLMs) such as CLIP, ViLBERT, and LXMERT implement this paradigm. For example, CLIP utilizes independent encoders for images and text, integrating their embeddings only at late stage via a contrastive alignment objective. Similarly, ViLBERT and LXMERT maintain distinct feature representations for image regions and language tokens before merging them through co-attentional or cross-modal transformer layers. While this delayed multimodal alignment ensures that each modality's features remain semantically coherent, it also requires additional fusion mechanisms—such as cross-attention modules or contrastive objectives—to unify these specialized embeddings into a shared, task-relevant latent space. Consequently, the late-fusion strategy introduces increased architectural complexity, computational overhead, and potentially longer inference and training times.

2) ***Unified Feature Spaces****:* Unified Feature Spaces (Shared Embeddings) represent a paradigm shift from earlier Vision-Language approaches that maintained distinct modality-specific representations toward architectures where visual and textual data share a common, modality-independent embedding space. Instead of separately encoding images and text and then aligning these representations at a later stage, unified methods embed both modalities into a single latent space from the outset, enabling a single model to handle multimodal inputs consistently without the need for specialized alignment or late-fusion.

One of the key concepts of Unified Feature Spaces is the representation of visual and textual modalities as token sequences.For example, images are decomposed into patch-level tokens and text is tokenized into subword units. By treating pixels and words as tokens from a shared, fixed token set, models such as PaLM-E and BLIP-2 can process all tokens through the same Transformer-based encoder architecture. This uniform tokenization pipeline eliminates the need for modality-specific processing, ensuring a consistent, modality-invariant representation space.Once visual and textual data are transformed into a common tokenized form, the model's multi-head self-attention and feed-forward layers operate over all tokens simultaneously. Without separate pipelines or dedicated fusion modules, the self-attention mechanism implicitly learns cross-modal relationships. The parameters and attention heads collectively identify patterns, correspondences, and correlations between image patches and textual tokens, allowing the model to align and integrate information from both modalities.This approach results in multimodal representations that capture cross-modal semantics, and by consolidating both modalities into a shared latent space from the outset, it simultaneously simplifies architectural complexity and produces more coherent and richly integrated multimodal embeddings.

3) ***Contrastive-Aligned Feature Spaces****:* Contrastive-Aligned Feature Spaces utilize contrastive learning to structurally arrange visual and textual embeddings within a unified latent representation.In this paradigm, the model optimizes a contrastive loss (e.g., InfoNCE or margin-based ranking loss) which brings semantically related image-text pairs (e.g., an image and its matching caption) closer together, while pushing apart mismatched pairs. By enforcing these relational constraints, the model's parameters encode semantic similarity directly into the spatial arrangement of embeddings.As training proceeds, conceptually related items—such as various depictions of a specific object and corresponding linguistic descriptors—naturally form localized clusters. These clusters then aggregate into larger, coherent latent spaces that capture high-level semantic structures. This topological clustering organization enables transitive alignment. For example, if image A is aligned with text A, and text A is semantically related to text B, then image A will reside in close proximity to text B as well. Such transitive relationships enhance global semantic consistency within the embedding space.

C. Vision-Language Integration Architectures

1) ***Pre-Transformer****:* Before the widespread adoption of Transformer-based models, vision-language integration architectures predominantly relied on convolutional neural networks (CNNs) for visual feature extraction and recurrent neural networks (RNNs) for linguistic encoding. In these frameworks, an input image was first processed through a CNN

pretrained on large-scale image classification datasets such as ImageNet to produce a fixed-dimensional vector representation capturing high-level visual semantics. Concurrently, textual inputs were transformed into continuous vector embeddings and subsequently passed to RNN-based encoders (e.g., LSTMs or GRUs) to model syntactic and semantic dependencies across text [30] [31] [32]. During this era, the integration of visual and linguistic modalities was largely limited to late-stage fusion operations, wherein the independently obtained feature representations were combined using relatively simple mechanisms such as concatenation, averaging, element-wise multiplication, or linear projections [33].The models inherently lacked fine-grained alignment strategies due to the absence of explicit mechanisms to focus attention on specific regions or objects within the visual input that corresponded to specific linguistic elements, resulting in images being treated at a global level, thereby limiting their ability to highlight salient objects or focus regions to specific linguistic elements [30].

Early fusion strategies introduced limited multimodal interaction by integrating visual and textual features at the initial stage of the decoding pipeline [30] [32]. For example, a global CNN-derived image embedding could be concatenated directly with the initial hidden state of an RNN-based language model, thereby placing both modalities into a shared representational space from the outset [32].As the fusion mechanism is fixed and non-adaptive, it does not enable dynamic adjusting of the relative contributions of each modality as the sequence progresses. As a result, patterns and features learned prematurely were reinforced, propagating low-level noise and incomplete representations throughout the model's layers. This static and non-adaptive early fusion approach constrained the complexity, accuracy, and quality of cross-modal grounding, limiting the model's ability to flexibly and effectively integrate information across vision and language.

Late fusion strategies delayed the integration of visual and linguistic modalities until both modalities' latent representations had been fully encoded. In this approach, the image encoder (e.g., a CNN) and the language encoder (e.g., an RNN or LSTM) independently extract high-level embeddings, which are subsequently combined only after each modality has achieved a semantically rich representation. This delay mitigated premature cross-modal interference, ensuring that each modality's embedding space has fully converged, normalized, and aligned prior to fusion, thereby enhancing the reliability and interpretability of the unified representation [31]. The lack of intermediate communication or joint attention mechanisms during encoding prevented the visual and textual encoders from influencing each other's internal representations, resulting in a unified representation that captures fewer fine-grained, context-sensitive relationships between the modalities. [30]

One of the key advancements in the pre-Transformer era was the integration of adaptive, sequential language models with CNN-based visual feature extractors. Unlike early or late fusion strategies—where linguistic representations remained fixed and were fused at a predetermined stage—recurrent architectures (e.g., LSTMs or GRUs) enabled dynamic updating of the language embedding throughout the decoding process. By incorporating CNN-derived visual embeddings at each time step, the model's hidden state could iteratively refine its predictions based on previously emitted tokens and evolving visual context [30] [32]. This iterative decoding paradigm facilitated the incremental integration of visual information, allowing lexical and syntactic decisions to be continuously adjusted as the sentence progressed. As a result, object-level grounding and fine-grained semantic alignment between vision and language were more effectively achieved. This adaptive, feedback-driven approach surpassed the limitations of static fusion methods, resulting in more contextually coherent and semantically integrated multimodal reasoning. [34]

This paradigm is exemplified by the Show and Tell [30] and Show, Attend and Tell [35] architectures, both of which introduce adaptive, sequence-aware language models to enhance the integration of visual and linguistic features. Show and Tell leverages a CNN pretrained on large-scale image recognition datasets (e.g., ImageNet) to encode the input image into a fixed-dimensional global embedding, effectively capturing high-level scene attribute information without explicitly identifying specific objects or regions. This global image representation is passed to an LSTM-based language decoder by either initializing the decoder's hidden state with the image embedding or concatenating it with the embedding of the first word token [30]. As the LSTM iteratively generates tokens, it dynamically conditions each subsequent output on its evolving hidden state—which captures previously emitted tokens and their linguistic dependencies—while continuously referencing the global visual embedding for contextual grounding. This approach outperforms static, single-step predictions by incrementally refining lexical and syntactic decisions, resulting in semantically coherent and contextually grounded captions.

Building upon these foundations, the Show, Attend and Tell [35] model introduces a key enhancement being a spatial attention mechanism. Rather than relying on a global image embedding, Show, Attend and Tell leverages CNN-encoded image features extracted at multiple spatial locations (e.g., feature maps from a convolutional layer) and learns to focus selectively on particular image regions at each timestep in the decoding process. This attention-based framework enables the LSTM decoder to focus on visually salient objects or attributes as needed, providing a more fine-grained and context-sensitive alignment between visual content and linguistic output. By dynamically shifting focus across different parts of the image while generating each word, the model achieves more precise object-level grounding and produces captions that are contextually consistent and visually descriptive. The incorporation of the attention mechanism represents a key technical development in adaptive, feedback-driven decoding being augmented

with fine-grained spatial awareness to achieve more semantically integrated multimodal reasoning. [35]

While earlier models, such as Show and Tell [30], generated coherent textual descriptions, they relied on global, image-level embeddings that constrained fine-grained vision-language alignment. This coarse representation limited the ability to associate specific linguistic elements with localized image regions, limiting detailed cross-modal grounding at the object or region scale. Deep Visual-Semantic Alignments for Generating Image Descriptions [31] addresses this limitation of global scene-level embeddings by introducing a more granular and context-sensitive approach to vision-language integration. Rather than encoding an entire image as a single, undifferentiated feature vector, the model utilizes region proposal methods (e.g., selective search) to isolate candidate bounding boxes, each capturing distinct visual objects or regions of interest. These object-level patches are then processed through a pretrained CNN (such as VGG16 or AlexNet) to produce region-centric feature embeddings, moving from a coarse global representation toward finer, localized visual descriptors.On the linguistic side, the approach similarly partitions textual input into semantically salient fragments—most notably noun phrases—identified through syntactic parsing tools (e.g., Stanford CoreNLP) [31]. These phrases are embedded into a continuous vector space (e.g., via word2vec), resulting in phrase-level textual embeddings that have the same level of granularity as compared to the object-level visual features. By representing both modalities as sets of discrete, semantically meaningful entities, the model enables a richer cross-modal mapping that surpasses simplistic, scene-level alignment. [31]

A key technical component is the projection of both region-level visual features and phrase-level textual embeddings into a shared multimodal embedding space, where the similarity between any visual region and textual phrase can be directly determined [31]. The model is trained using a structured max-margin loss ranking objective designed to increase the similarity between correctly paired region-phrase matches while decreasing it for mismatched pairs. This optimization shapes a topologically meaningful embedding space in which related visual and linguistic elements form coherent clusters, facilitating more robust and interpretable vision-language alignments.By integrating object-level visual embeddings, phrase-level textual representations, and a discriminative max-margin ranking objective, this approach achieves more precise object-level grounding and semantic integration.

*2) **Two-Stream (Multi-Stream) Transformer**:* Building upon the pre-Transformer era's developments in fine-grained alignment and iterative, context-aware decoding, vision-language integration architectures evolved toward Transformer-based frameworks that more explicitly separate unimodal encoding from multimodal fusion.Rather than using a single model to process both visual and textual data jointly, these architectures introduced Two-Stream (or Multi-Stream) Transformer frameworks to leverage modality-specific encoding pipelines. [18] In this configuration, each modality—visual and textual—is initially handled by a dedicated Transformer encoder stack, allowing each stream to undergo self-attention and feed-forward transformations specific to its unique representational structure.

[4] [5] The visual stream utilizes pre-trained object detectors (e.g., Faster R-CNN) operating on large-scale image datasets generate region-level feature embeddings that highlight salient objects or areas within the scene. These embeddings are refined by a Transformer-based visual encoder, which leverages multi-head self-attention, feed-forward transformations, and layer normalization to capture intra-visual relationships such as object co-occurrence pattern, spatial configuration, and attribute distribution information at a granular level. Similarly, the language stream utilizes a Transformer-based text encoder (e.g., a BERT-style architecture) to produce contextualized token embeddings that model syntactic and semantic dependencies across the input text. This unimodal separation ensures that each stream's encoder stack is focused on modality domain-specific abstractions unconstrained by the noise or structural biases of the other modality. [36] [37] After both streams have independently produced semantically rich, contextualized representations, a dedicated cross-modal attention or co-attention module fuses the information, establishing direct correspondences between visual regions and linguistic tokens. By dynamically adjusting attention weights, the model identifies and aligns semantically related elements from the two modalities, enhancing multimodal reasoning and improving semantic grounding.

[4] ViLBERT represents a significant advancement in two-stream (multi-stream) Transformer-based Vision-Language (VL) integration. Adapting BERT's representation learning, ViLBERT jointly models visual and textual modalities using parallel Transformer encoder stacks. By deferring multimodal fusion until after unimodal representations have been learned, ViLBERT facilitates more accurate and fine-grained semantic alignment across modalities.

[4] In ViLBERT, the visual stream utilizes a pretrained object detector (e.g., Faster R-CNN) to isolate and encode region-level features such as bounding boxes and associated embeddings that capture object appearance, attribute, and spatial relational information. These object-centric embeddings serve as visual tokens and are refined by a Transformer-based visual encoder utilizing multi-head self-attention and feed-forward layers. Through this process, the model contextualizes each region-level embedding, capturing global scene context and complex inter-object dependencies.Simultaneously, the linguistic stream processes tokenized textual inputs enriched by token, positional, and segment embeddings. A BERT-like Transformer encoder applies multi-head self-attention and feed-forward transformations to generate contextually enriched

token embeddings, effectively modeling syntactic structure, semantic content, and lexical relationships within the text. By independently refining visual and linguistic inputs, ViLBERT ensures that each modality is internally coherent and semantically rich before cross-modal interaction.

[4] A key architectural innovation is ViLBERT's approach to iterative multimodal fusion. Rather than confining cross-modal integration to the input or output layers, ViLBERT interleaves the unimodal self-attention layers with co-attentional Transformer layers that implement bidirectional cross-attention. [37] In ViLBERT's co-attentional layers, attention operates bidirectionally between the visual and linguistic streams. During vision-to-language (V→L) attention, visual embeddings function as queries, while linguistic embeddings serve as keys and values. The model utilizes the visual representations as reference points to determine which linguistic elements are most relevant, thereby enriching the visual latent space with semantic information derived from the text. Conversely, during language-to-vision (L→V) attention, linguistic embeddings act as queries, and visual embeddings provide keys and values. In this phase, the textual representation selects salient image regions to ground the semantic concepts expressed by the words.Through repeated rounds of V→L and L→V attention, each modality iteratively influences the other, adjusting and refining its representations based on received complementary information. This iterative, bidirectional querying process aligns and enhances both the visual and linguistic embeddings, resulting in a mutually reinforced contextually rich and semantically precise multimodal representation better integrated across the vision and language modalities.

[5] LXMERT utilizes a two-stream, Transformer-based framework that explicitly segments unimodal representation learning from subsequent multimodal integration. This approach differs from ViLBERT, which interleaves cross-modal interactions at earlier stages of the encoding process. LXMERT ensures robust, domain-specialized embeddings in both linguistic and visual streams by deferring multimodal fusion until after the unimodal encoders have fully refined their respective representations. LXMERT can be considered as a complementary advancement to ViLBERT as both models introduce similar two-stream, Transformer-based architectures for vision-language integration, but they implement different strategies for integrating the modalities. ViLBERT interleaves unimodal and cross-modal attention early, aiming for continuous mutual influence throughout the encoding process. LXMERT, postpones cross-modal fusion until each modality has been independently encoded resulting in a more modular and segmented multimodal integration.

[5] For the linguistic stream, LXMERT's language encoder follows a BERT-like Transformer architecture. It processes tokenized input augmented with token, positional, and segment embeddings, using multi-head self-attention and feed-forward layers to capture syntactic dependencies and semantic relations among words. In parallel, the visual encoder refines region-level features derived from a pretrained object detector. Through Transformer layers, the visual stream establishes meaningful scene-level context by modeling spatial and semantic relationships between proposed objects or regions. This unimodal processing phase for vision and language streams, ensures that each modality attains a rich, domain-specialized representation.

[5] After independently encoding each modality LXMERT's cross-modality encoder integrates these unimodally refined features using co-attentional Transformer layers that implement bidirectional cross-attention. Linguistic embeddings query visual features to ground abstract textual concepts in image regions, while visual embeddings reference linguistic tokens to reinforce semantic clarity. By iteratively applying bidirectional query-key-value mechanism co-attentional operations, LXMERT incrementally aligns both modalities within a unified embedding space, resulting in context-rich, and semantically integrated vision-language representations.

3) ***Single-Stream (Unified) Transformer****:* Single-Stream (Unified) Transformer architectures mark a developmental step in Vision-Language (VL) modeling by eliminating the separation of unimodal encoding and delayed fusion. In these unified frameworks, the model processes both visual and textual inputs simultaneously into a single Transformer-based encoder, allowing them to interact continuously at every layer. Instead of having distinct pipelines for vision and language, unified approaches interleave image-derived features (e.g., region-level embeddings or patches extracted by a CNN or a Vision Transformer) with textual tokens in a shared input sequence.

A key technical feature of unified architectures is the application of self-attention over a joint set of multimodal tokens. Each Transformer layer considers both linguistic tokens and visual elements as part of the same input, enabling the attention heads to discover fine-grained correspondences and dynamically align words, phrases, and object-level features at multiple abstraction levels. This design contrasts with two-stream models that first produce unimodal representations and only later integrate them through a separate co-attentional mechanism. Early and continuous modality interaction allow unified Transformers to more efficiently learn visual-linguistic representations, improving semantic grounding, object-level alignment, and context-sensitive reasoning. [38]

[14] VisualBERT is an key development in single-stream (unified) Transformer-based architecture for vision-language integration, as it integrates visual and textual inputs from the outset instead of initial unimodal representation derivation and subsequent fusion.VisualBERT's single-stream architecture integrates textual tokens and region-level visual embeddings into

a unified sequence before passing them into a Transformer-based encoder, enabling the self-attention mechanism to model visual-linguistic relationships continuously at every layer.By placing both text and image representations in the same input sequence, each embedding—whether it originates from language or vision—is processed through the same stack of Transformer layers [37]. Within each layer's multi-head self-attention, queries, keys, and values are derived from this joint set of multimodal tokens, allowing every token to directly attend to relevant elements from both modalities.

[14] As the model progresses through successive layers, the embeddings are iteratively updated, and cross-modal correspondences are refined.For example, a linguistic token representing a noun can attend to a corresponding image region to clarify its visual semantics, while visual embeddings derived from detected objects can attend to words or phrases that provide context and meaning. This iterative refinement negates the need for separate unimodal pipelines or late-stage fusion steps, as every layer inherently performs multimodal integration through shared self-attention.Each token's representation encodes a rich, contextually informed blend of linguistic and visual information. This continuous, layer-by-layer alignment encourages fine-grained semantic grounding and enhances context-sensitive multimodal reasoning.

[17] UNITER extends the single-stream (unified) Transformer-based vision-language modeling paradigm introduced by models such as VisualBERT [14] by incorporating explicit spatial information and refining its multimodal embedding strategy.UNITER jointly processes textual tokens and region-level image embeddings—extracted by a pretrained object detector within a single Transformer encoder, allowing continuous cross-modal interaction at every layer.While VisualBERT treats each detected object region as an isolated feature vector appended to a text-centric framework, UNITER augments these visual embeddings with continuous bounding-box coordinates and integrates them directly alongside textual tokens into a unified input sequence. This architecture establishes a coherent multimodal embedding space from the outset, enabling the Transformer's self-attention mechanism to consider both linguistic elements and spatially grounded visual features simultaneously.

A key architectural improvement is in the encoding of visual information.While VisualBERT treats each detected object region as an isolated feature vector, UNITER augments visual embeddings with continuous bounding-box coordinates, providing the Transformer's attention mechanism with explicit spatial references for each detected object region. As a result, the model can interpret what objects are present, where they are located and how they relate to one another in the scene. [17]

[17] Rather than treating objects as isolated feature vectors, UNITER integrates both the semantic and spatial dimensions of visual information into a unified embedding space. By incorporating bounding-box coordinates, the model can consider not what objects are present and also where they are located and how they relate to one another in the scene [33]. For example, in an image captioned as "A man in a red jacket standing on skis next to a black dog," UNITER receives embeddings for the detected objects—man, red jacket, skis, black dog—along with their bounding-box coordinates. This allows the Transformer to align the textual references precisely with the corresponding image regions. For instance, the token "man" and the phrase "red jacket" are associated with the correct object embedding and the bounding-box that pinpoints the man wearing that jacket [32]. Similarly, the phrase "next to" becomes spatially grounded, enabling the model to focus attention to the relative position of the dog's bounding box in relation to the man.

[17] As UNITER arranges visual features, spatial coordinates, and textual embeddings into a single multimodal sequence, the Transformer's self-attention mechanism treats all inputs—regardless of their modality—as part of one coherent set of tokens. This unified format ensures that no single modality is prioritized or processed in isolation. Instead, the attention heads dynamically weigh and align relevant linguistic and visual elements at every encoding layer, drawing connections between words and their corresponding image regions while also considering the spatial relationships encoded in bounding-box coordinates [33]. Queries, keys, and values are derived from this unified set of tokens, allowing the attention heads to naturally weigh elements from both modalities and refine alignments at every encoding layer [36]. A textual token such as "man" can issue queries that highlight the visual embeddings and bounding-box coordinates corresponding to that concept, while visual tokens can similarly attend to linguistic elements that provide semantic context or clarify object attributes [38]. By eliminating the need for separate or delayed fusion steps, the model supports direct and continuous cross-modal interactions [32]. This integrated approach enables the Transformer's self-attention layers to iteratively discover and strengthen meaningful correspondences between language and vision, producing richer, more semantically grounded multimodal representations that enhance vision-language performance.

[20] ViLT presents a streamlined, single-stream (unified) vision-language (VL) modeling paradigm that removes the reliance on region-level object detectors common in earlier models by entirely avoiding region-level object detection and instead implementing a patch-based embedding strategy inspired by the Vision Transformer (ViT).Rather than extracting object-centric embeddings through a pretrained detector—an approach that introduces complexity, fixed vocabularies, and preprocessing overhead [23]—ViLT divides input images into grids of patches and linearly projects each patch into an embedding vector, effectively treating them as visual tokens [7], enabling the model to learn directly from raw, domain-

independent pixel-level data [27], promoting a more generic and scalable representation.

[20] ViLT streamlines and improves vision-language integration by eliminating the need for region-level object detection and adopting a patch-based approach inspired by the Vision Transformer (ViT) [25] [3]. Instead of extracting features from predefined object regions, ViLT partitions images into a grid of patches, each treated as a visual token, and processes these alongside textual tokens within a single Transformer encoder. This design frees the model from reliance on fixed object vocabularies and object detectors, reducing architectural complexity and overhead. By integrating visual and textual tokens into a unified multimodal sequence [27], ViLT ensures continuous, layer-by-layer interaction between modalities, allowing words to attend to spatially grounded patches and image features to incorporate semantic information from text at every encoding stage. [37] Over successive layers, this iterative, bidirectional exchange produces richer, more context-sensitive, and semantically integrated multimodal representations.

4) ***Decoupled Vision-Language Encoders with Cross-Attention Layers****:* Decoupled vision-language encoders with cross-attention layers represent a recent architectural paradigm that aims to balance unimodal specialization and efficient multimodal integration. Each modality—vision and language—is first processed by its own dedicated encoder,a pretrained transformer model such as a ViT or CLIP encoder for images, and a Large Language Model (LLM) for text, ensuring that both modality streams produce high-quality, domain-specialized representations independently and without immediate cross-modal influence.Once these unimodal representations are extracted, a lightweight cross-attention interface selectively integrates the two streams [37]. Rather than merging the encoders into a single model or employing deeply interwoven co-attentional blocks, this decoupled approach introduces a small number of cross-attention layers to link the visual and linguistic representations, injecting just enough cross-modal parameters to align and enhance the combined embedding space. [27]

This architectural approach preserves the scalability and modularity of unimodal models by allowing both the vision and language encoders to be independently improved, updated, or replaced without reengineering the entire multimodal pipeline. Additionally, utilizing pretrained encoders ensures that each modality's representation space is well structured and semantically rich before cross-modal integration, which can lead to more effective and stable multimodal alignment. By decoupling the unimodal encoders and introducing only a minimal cross-attention interface, the approach streamlines vision-language integration into a more modular process.This reduced complexity enables the model to adapt more readily to improved unimodal components, resulting in more flexible and scalable vision-language integration approach.

[39] BLIP-2 exemplifies how vision and language models can be integrated efficiently into Large Language Models (LLMs) through a decoupled architecture coupled with minimal cross-attention layers. Rather than merging the vision and language components into a single unified pipeline, BLIP-2 leverages large, independently pretrained unimodal encoders—such as a frozen vision encoder (e.g., ViT or CLIP) and a state-of-the-art LLM like LLaMA or Flan-T5—each already optimized within its respective domain.

[39] BLIP-2 exemplifies how vision and language models can be integrated efficiently into Large Language Models (LLMs) through a decoupled architecture coupled with minimal cross-attention layers. Rather than merging the vision and language components into a single unified pipeline, BLIP-2 leverages large, independently pretrained unimodal encoders—such as a frozen vision encoder (e.g., ViT or CLIP) and a state-of-the-art LLM like LLaMA or Flan-T5—each already optimized within its respective domain.

BLIP-2's approach utilizes the Q-Former, a small, dedicated Transformer module that provides a minimal, learnable interface for vision-language integration in LLMs.Instead of tightly coupling the two modalities, the Q-Former translates the vision encoder's output embeddings into query tokens that the LLM can readily interpret as input text. This approach allows the LLM to incorporate visual context—such as scenes, objects, and their relationships—directly into its reasoning process, all without the need for large-scale multimodal retraining. In doing so, BLIP-2 augments the LLM's language-based reasoning capabilities with visual grounding, enabling it to generate richer, more context-aware outputs that reflect both linguistic and visual information.By decoupling the modalities and confining cross-modal fusion to a lightweight Q-Former module, BLIP-2 achieves parameter-efficient, dynamic multimodal reasoning.BLIP-2 effectively expands the capabilities of LLMs by integrating robust visual understanding, leading to more comprehensive and semantically grounded multimodal reasoning.

III. Vision-Language Models Training

In the evolution of deep learning, Bidirectional Encoder Representations from Transformers (BERT) [40], a Transformer-based [18] architecture, has distinguished itself in the field of language modeling by significantly outperforming many of its contemporaries and has led researchers to explore extending its capabilities to visual data processing. The success of BERT has led researchers to explore extending its capabilities to visual data processing, leading to cutting-edge models such as visual-BERT [14] and ViLBERT [4], which incorporate textual and image markup to achieve even greater power. The model training strategy focuses on two key objectives. One task is mask modeling. This task requires the model to predict the content of partially masked elements in the input based on context. The model must also investigate

the underlying patterns and logical associations within the data. The second is the sentence-image prediction task, in which the model needs to accurately determine whether a given subtitle accurately portrays the essence of the image to strengthen the ability to understand and grasp the consistency between visual and textual semantics. With the synergy of these dual goals and the unique attention mechanism of the transformers architecture, which is capable of capturing the subtle correlations between words and visual cues in complex data, the model has demonstrated excellent performance in many key tasks of visual-linguistic fusion, laying a solid foundation and pioneering the research of visual-linguistic modeling in the future. The model has demonstrated excellent performance in many key tasks of visual-linguistic fusion processing, laying a solid foundation for subsequent visual-linguistic modeling research and opening up new directions. With the impressive development of deep learning in computer vision and natural language processing, the Meta team [41] has subdivided the latest Transformer-based techniques into four training paradigms. Notably, these paradigms are not mutually exclusive, with many approaches relying on a mix of contrastive, masking, and generative criteria.

A. Contrastive-based VLMs

Contrastive learning is crucial in VLMs as a practical framework for cross-modal alignment. VLMs based on contrastive learning typically consist of two main components: a visual encoder and a text encoder. The visual encoder, often built on convolutional neural networks (CNNs) or visual transformers (ViTs) [3], extracts high-dimensional embeddings from input images. The text encoder, in contrast, utilizes pre-trained language models to convert textual inputs into semantic embeddings. Once these embeddings are projected into a shared high-dimensional space, they are processed by a learnable projection head to ensure compatibility between the visual and textual representations. This alignment supports downstream tasks such as retrieval, classification, and generation.

The central idea of contrastive learning is to efficiently align visual and textual modalities by maximizing the similarity of positive sample pairs and minimizing the similarity of negative sample pairs in a shared latent space. Positive sample pairs are mapped closer in the latent space; while negative sample pairs are mapped farther away. The model is trained to differentiate between similar and dissimilar samples. To explain this process, Energy-Based Models (EBMs) [42]serve as a theoretical framework. In EBMs, the goal is to assign low energy to observed data and high energy to unobserved data. Specifically, the true data distribution should have low energy, while noise or other irrelevant data points should have high energy. This is achieved by representing the ”energy” of input data using an energy function and training the model to match the data's energy distribution with the expected distribution. Several methods, such as Markov Chain Monte Carlo (MCMC), have been proposed to approximate this distribution during training. MCMC iteratively finds negative samples from the model distribution, minimizing predicted energy. However, direct sampling from the model distribution is infeasible, which led to the development of Noise Contrastive Estimation (NCE) [43]. NCE approximates model sampling by selecting negative samples from a noise distribution, simplifying training, and avoiding complex normalization factors. It is widely used in models like Word2Vec [44], where negative samples are sampled from noisy vocabularies. NCE improves alignment by optimizing the model's ability to distinguish between positive and negative sample pairs. In the InfoNCE [45] framework, the model further refines alignment by calculating the cosine similarity between embeddings and predicting the most likely positive pairs through Softmax while assigning lower probability to negative pairs. A temperature parameter in the loss function regulates the sensitivity of the alignment, balancing the model's ability to distinguish subtle semantic differences.

Numerous landmark models combine contrastive learning with dense alignment techniques, achieving superior performance on visual language tasks. Traditional contrastive learning focuses on instance-level alignment, optimizing global features by bringing representations of the same instance closer together and separating representations of different instances. SimCLR [46], for example, enhances the global feature representations of images by maximizing the distances between different samples while minimizing distances between similar ones. It achieves this by augmenting images (e.g., through cropping, rotation, or color changes) and treating these augmented images as different views of the same instance for comparative learning. MoCo [47] also optimizes global feature alignment using a dynamic dictionary to store feature representations of different instances, incorporating Momentum updating to maintain stability during dictionary updates. Contrastive Language-Image Pretraining(CLIP) [1], developed by OpenAI, utilizes 400 million image-text pairs and trains both visual and text encoders through contrast loss. It treats each image-text pair as a positive sample while considering all other combinations in the batch as negative. CLIP excels in zero-shot learning tasks, using natural language cues for classification without task-specific fine-tuning. Building on CLIP, A Large-scale Image and Noisy-text embedding (ALIGN) [7] extends instance-level alignment by utilizing 1.8 billion noisy image-text pairs, demonstrating the robustness of contrastive learning with noisy data. The embeddings generated by ALIGN perform well in downstream tasks like retrieval and classification.

These methods emphasize instance-level alignment and learn overall global features, paying more attention to the overall semantic information of the data. However, new technologies such as Dense Image-Text Alignment [48] retain local details or fine-grained feature expressions and expand this paradigm by integrating fine-grained correspondences between image regions and text elements. This alignment method

represents images as region-level embeddings and texts as word or subword embeddings. The model no longer only aligns overall embeddings but optimizes region-word or token-token similarity to achieve finer alignment. The Florence [49] model proposed by Microsoft is a typical representative of this technology. Florence achieves dense alignment by calculating the attention-weighted similarity between visual patches and text tokens. This mechanism helps the model identify specific regions in an image and establish connections with relevant words in the description. The dense alignment strategy significantly improves performance in object-level retrieval, visual question answering, and scene understanding tasks.

The impact of contrastive-based VLMs extends beyond traditional image-text retrieval and text generation tasks. These models perform exceptionally well in zero-shot classification. They require no task-specific fine-tuning to achieve effective domain adaptation.

B. VLMs with Masking Objectives

Masking techniques have an important place in the field of deep learning, especially in VLMs, where masking goals are widely used to enhance the cross-modal learning ability of the models. Masking techniques first originated from Denoising Autoencoder (DAE) [50], which learns a latent representation of the data by adding noise to the input data and training the model to reconstruct the original data. In this framework, the noise has a spatial structure, which is usually achieved by masking parts of the data randomly. Masking techniques are closely related to image inpainting strategies, where the goal of image restoration is to recover missing or damaged parts of an image. The approach proposed by Pathak et al. [51] in 2016 utilizes this idea by learning a strong representation of an image to handle the image restoration task. In the field of language processing, Masked Language Modeling (MLM) also has a similar idea, which learns the underlying structure and grammatical rules of a language by masking certain words in a sentence and training the model to predict these missing words. BERT is a typical example of the adoption of the MLM strategy, which significantly improves the natural language processing task by randomly masking some of the words and predicting them at training time, which significantly improves the performance of natural language processing tasks.

In recent years, masking targets have been gradually introduced into VLMs, facilitating the progress of multi-modal learning. In VLMs, masked targeting, as a self-supervised learning method, can effectively deal with the relationship between different modalities, especially in the task of fusing text and image information. The basic principle of masked targeting is to learn a stronger cross-modal representation by masking a portion of the input and forcing the model to infer the hidden information from the remaining portion. The application of masking targets helps to create a stronger link between vision and language, improving the performance of the model in complex tasks.

FLAVA [52] is a typical masked target-based VLM that processes visual and textual data through the synergy of three core components. FLAVA employs a ViT as an image encoder that slices the input image into multiple chunks and linearly embeds them, which are then processed through the transformer architecture. The text encoder, on the other hand, tokenizes the text input, embeds it as vectors through the standard transformer architecture, and outputs hidden state vectors. FLAVA combines both multi-modal and unimodal masking modeling losses during training and incorporates a comparative learning objective to achieve leading scores on a number of visual, linguistic, and multi-modal tasks. Oscar [53] is also a successful VLM that combines self-supervised learning with visual-language masking goals. Oscar uses "object-level" masking, i.e., it allows the model to learn how to comprehend and align objects in multimodal data by masking out certain object regions in the image and hiding the vocabulary associated with these objects in the text. In this way, Oscar enhances the link between vision and language, further improving the performance of visual-text tasks. In the Oscar model, the masking strategy is not limited to randomly selecting part of the information in a text or image, but by strategically selecting and masking certain objects, it helps the model to better understand the deeper associations between visual and verbal information.

The introduction of masking objectives makes the training of VLMs more efficient and flexible. Through self-supervised learning, the masking strategy not only helps the model to extract more profound inter-modal relationships but also effectively improves the model's generalization ability in multi-modal tasks. This approach provides new ideas for the integration of vision and language and lays a theoretical foundation for future multi-modal learning.

C. Generative-based VLMs

Generative-based VLMs have made significant progress in recent years, especially in text-to-image, image-to-text, and cross-modal understanding and generation. Unlike traditional discriminative models, the core strength of generative models lies in their ability to understand and create entirely new data samples, which makes them show great potential in tasks such as image description generation, visual quizzing, and image content generation. By jointly modeling the distribution of image and text data, the generative model is able to deeply capture and parse the complex and diverse connections between images and language. This joint learning approach not only allows the model to have a more precise understanding of the relationship between the two modalities, but also demonstrates stronger generalization ability in cross-modal tasks, thus significantly improving the overall performance. This generative capability not only enhances the models'

flexibility in visual language tasks, but also allows them to show greater adaptability and creativity when dealing with complex multimodal data.

A typical generative visual language model is DALL·E [54], proposed by OpenAI, which is capable of generating images associated with textual inputs. DALL·E is based on the Transformer architecture. It learns deep connections between image and textual data through co-training. This approach allows DALL·E to generate semantically accurate and creative images. It demonstrates the effectiveness of generative models in visual-verbal tasks. In addition, the Imagen [55] model employs a similar generative model architecture, but further improves the quality of the generated images by introducing techniques such as reinforcement learning. In particular, Imagen outperforms DALL·E in the reproduction of details and textures and is able to generate more detailed and realistic images, thus performing even better in generative tasks.

Similar to DALL·E and Imagen, Google's Parti [56] model focuses on generating high-quality images, especially in the missing details of the textual descriptions. Parti avoids the problem of blurry or unclear images that are common in traditional generative models by progressively refining the image-generation process. Parti's uniqueness lies in its latent space-based generation methodology, which gives the model greater flexibility to efficiently handle generation tasks for different styles and scenes. This innovation enables Parti to excel in diverse generation tasks. It significantly improves the quality of generated images, resulting in more refined and diverse visual outcomes.

In addition to the above models, CoCa [57] is also a model that has achieved remarkable success in the field of generative-based VLMs. CoCa combines the strengths of contrastive learning and generative learning by utilizing a large amount of unlabeled data for pre-training and thus efficiently learns the deep associations between images and text. The core idea of CoCa is to optimize the text-image alignment by using the loss of contrasts, which enables it to generate image-text pairs with a high degree of consistency. In the image generation task, CoCa generates visually appealing images. It also ensures that these images closely match the input text descriptions. As a result, CoCa demonstrates excellent performance in generation accuracy.

As technology continues to advance, the potential of generative-based VLMs will be realized in more practical applications. In the future, generative visual language models may not only be able to generate high-quality images and text but also be able to understand complex cross-modal tasks, thus driving the further development of multimodal AI systems. The wide application and continuous optimization of these models will help to address the challenges in visual-linguistic integration better and drive the progress and innovation of intelligent systems in multiple domains.

D. VLMs from Pre-trained Backbones

Building VLMs from scratch is undoubtedly challenging, with the core difficulties of high cost and inefficient computation [41]. First of all, if a VLM is trained from scratch, two distinct modalities, visual and linguistic, must be mastered simultaneously, meaning that not only image feature extraction but also linguistic semantics modeling must be handled. For a model to have good generalization ability, it often needs to rely on large-scale cross-modal datasets, and labeling such data is undoubtedly a time-consuming and labor-intensive task. In addition, the combination of multiple modalities makes the number of parameters in a multimodal model far exceed that of a single-modality model, and the high consumption of computational resources for training is a huge burden for researchers and enterprises. To cope with this dilemma, VLMs from Pre-trained Backbones have emerged to provide a more efficient and economical alternative. Pre-trained models in the visual and linguistic domains have been deeply trained on massive datasets, capturing and refining the generic feature representations of their respective modalities. These pre-trained weights can be used as the basis for visual language models, and excellent model performance can be quickly achieved through migration learning and a small amount of cross-modal task fine-tuning, greatly improving training efficiency.

LLaVA [58] is a cross-modal dialog AI assistant that combines LLaMA [59] and CLIP, focusing on image and text synergistic understanding.LLava utilizes GPT-4 to convert image-text pairs into instruction-following formats, thus collecting multimodal instruction data covering three categories: dialog, detailed description, and complex reasoning, with 158K samples. The model architecture connects the CLIP visual encoder with the Vicuna language decoder, which is trained in two stages of instruction tuning. In the first stage, we pre-train the alignment features, filter and transform some image-text pairs of CC3M, and train only the projection matrix; in the second stage, we update the weights of the projection layer and the language model according to different task scenarios in order to improve the model's visual instruction-following and reasoning capabilities.

BLIP-2 [39] innovatively proposes a visual language pre-training method, which is centered on the clever use of frozen pre-trained image encoders and LLMs, the use of a lightweight Q-Former to bridge the modal effectively, so as to achieve the performance improvement and cost reduction through the well-designed two-phase pre-training strategy. At the model architecture and pre-training target level, Q-Former covers image and text transformer sub-modules with a total of 188M parameters. For the pre-training setup and its significant advantages, a rich dataset combined with the CapFilt [2] method to generate synthetic subtitles is used as the data support, the ViT model is carefully selected to be the image encoder, and OPT [60] or FlanT5 [61] is used as the LLM so that the pre-training process can be carried out in phases

and in an orderly manner under specific parameter conditions. With this unique approach, BLIP-2 can demonstrate excellent performance in visual-linguistic tasks with a small number of trainable parameters and is capable of zero-sample instruction image-to-text generation.

IV. Evaluation

A. VLM Benchmarking

Vision language model bridge the gap between visual and text information, enabling large models to be used in broader scenarios beyond text generation, translation, etc. VLM benchmarking refers to using evaluation datasets and metrics to test the performance of VLM models. At the very early stage of VLM development, evaluation for VLM is divided into two tasks, image captioning and Visual Question Answering.

B. Image Captioning

Image captioning is the task of automatically generating descriptive textual captions for an image. This task is essential for evaluating the ability of VLMs because it directly assesses the ability to understand images and generate text. Image captioning typically involves two parts, the first part is an encoder for image processing and feature extracting, and the second part is a decoder, taking in the extracted features and generating text word by word. Both language ability and vision ability for VLMs are crucial in performing image captioning tasks. The COCO [62] (Common Objects in Context) dataset is widely used in the evaluation of image captioning. COCO is a large-scale dataset that is designed for training and testing in object detection, segmentation, and captioning. It has 330K images and more than 200K of them are labeled. COCO also has a tailored feature for image captioning tasks, as it provides 5 captions per image in the captioning subset. The captioning subset has 413,915 captions for 82,783 images in training, 202,520 captions for 40,504 images in validation, and 379,249 captions for 40,775 images in testing. Evaluation of performance using the COCO dataset in image captioning involves using third-party metrics [63], including BLEU [64], METEOR [65], ROUGE-L [66], and CIDEr [67]. These metrics measure the alignment between generated captions and true captions. However, these metrics do not truly reflect the captioning performance due to the existence of false negative issues. Since there are only limited and determined captions for each image, applying these metrics will reduce the score for similar captioning that varies in expressions. Many recent works introduce large modality models into the evaluation process, which make the evaluation more robust by mitigating false negative problems.

C. Visual Question Answering

Visual Question Answering is a more challenging and comprehensive task compared to image captioning. VQA involves answering diverse, context-rich questions about images, making it a challenging task that tests both visual recognition and linguistic reasoning. Popular datasets like the VQA v2 [68] dataset provide a foundational benchmark, with its extensive question-answer pairs from COCO images. CLEVR [69] focuses on compositional reasoning using synthetic scenes. Other datasets like GQA [70] and VizWiz [71] push models toward improved reasoning and robustness, covering structured scene graphs and real-world scenarios, respectively.

More challenging VQA tasks do not limit testing VLM ability in simply processing images and text. Many other tasks also involve the ability to use general knowledge for analysis, logical reasoning, etc. OK-VQA [72] (Outside Knowledge VQA) is a challenging Visual Question Answering dataset designed to test models' ability to answer questions requiring external knowledge beyond the image content. Unlike standard VQA datasets, where questions are mostly answerable by analyzing visual information, OK-VQA includes open-domain questions that demand reasoning and factual information sourced from general knowledge. For example, a question might ask, "What type of animal is shown in this picture?" where recognizing a specific breed or habitat would require prior knowledge. The dataset includes approximately 14,000 images and 30,000 questions derived from the COCO dataset, encouraging the development of systems that combine visual understanding with knowledge retrieval mechanisms to bridge the gap between perception and reasoning.

As the reasoning ability of large models remains questioned by the general public, benchmarks designed from challenging tasks in mathematical reasoning from visual contexts become a popular way to test the reasoning ability for VLMs.

The MATHVISTA [73] dataset is a comprehensive benchmark designed to evaluate mathematical reasoning in visual contexts. It addresses a critical gap in existing datasets that primarily focus on textual problems. It includes 6,141 examples derived from 28 multimodal datasets and three newly created datasets—IQTest, FunctionQA, and PaperQA. These tasks span diverse mathematical reasoning types, such as algebraic, geometric, logical, and statistical reasoning, across various visual contexts like natural images, charts, function plots, and diagrams. MATHVISTA's examples are annotated with metadata, including question type, grade level, and reasoning skills, and are curated to challenge AI models with fine-grained visual understanding and compositional reasoning.

The MATHVERSE [74] dataset is a benchmark for evaluating the multi-modal mathematical reasoning skills of VLMs. It contains 2,612 visual math problems drawn from plane geometry, solid geometry, and functions, transformed into six

distinct versions, contributing to over 15,000 test instances. These versions systematically adjust the textual and visual content to assess whether models can truly interpret diagrams for mathematical reasoning rather than relying solely on textual descriptions. Additionally, MATHVERSE introduces a novel Chain-of-Thought (CoT) evaluation strategy, leveraging GPT-4 to assess step-by-step reasoning, enabling detailed error analysis and fine-grained scoring. This benchmark aims to address limitations in current datasets by challenging models to balance textual and visual reasoning for solving math problems.

D. Benchmarking Hallucination

Hallucinations are a significant challenge for large language models as well as vision language models. These models often generate information with high confidence that appears plausible but is entirely false. Similarly, vision-language models (VLMs) could hallucinate by generating text or captions unrelated to the images they are tasked with describing. This is unacceptable and significantly harmful when VLMs are used in scenarios where a single hallucination or mistake will drive bad results. For example, in the self-driving scenario, if the VLM in the system hallucinates and provides wrong commands to the driver, it may cause severe safety problems. That is the reason why in most autonomous driving systems, the end-to-end model is the one that really does the self-driving part, where VLMs take into place when encountering more complicated road conditions and provide only suggestions for the human driver. It is still not safe to trust VLMs in many real-world applications. Therefore, evaluating whether VLMs produce hallucination-free outputs is a critical area of research.

CHAIR [75] is the first benchmark for detecting object hallucinations in image captions. It focuses on a fixed set of objects from the COCO dataset and is especially useful for evaluating short, single-sentence captions. However, CHAIR has limitations when applied to longer outputs from modern VLMs. It may misclassify hypothetical statements as hallucinations, and overlook hallucinations outside the predefined object set. And given its limitation in the COCO dataset, which is an almost must-have training set for modern VLMs, CHAIR lacks the ability to do a comprehensive evaluation of hallucination.

Recent research, POPE [76], offers a more flexible approach by using binary polling questions, including both positive (grounded in true objects) and negative (derived from unrelated objects). GAVIE [77] (GPT4-Assisted Visual Instruction Evaluation) leverages the power of using GPT-4 to generate visual instructions, and uses LRV-Instruction includes both positive and negative instructions, designed across three semantic levels: Nonexistent Object Manipulation, Existent Object Manipulation, and Knowledge Manipulation, to enhance robustness.

Most recent hallucination benchmark, which is also used by HuggingFace OpenVLM Leaderboard, is called Hallusion-Bench [78]. This benchmark is designed to evaluate advanced large visual-language models (LVLMs) like GPT-4V (vision), Gemini Pro Vision, Claude 3, and LLaVA-1.5 on nuanced image-context reasoning. It has 346 images and 1129 expert-crafted questions, emphasizing logical consistency, response tendencies, and failure modes through a novel question structure enabling controlled analyses. The same team recently develop a new framework to automatically generate benchmark for studying hallucinations in large vision-language models (LVLMs). Traditional benchmarks rely on hand-crafted corner cases with limited generalizability, where the new method, AUTOHALLUSION [79] scales up the creation of hallucination examples through an automated approach, minimizing human bias. This framework will benefit researchers in constructing benchmark dataset and could be used together with human-crafted cases.

E. Comprehensive Benchmarks

A holistic benchmark like MMBench [80] for vision-language tasks is becoming more popular than traditional benchmarks because it evaluates models across a broader spectrum of tasks, datasets, and metrics, offering a more complete picture of their strengths and weaknesses. Unlike narrow benchmarks that focus solely on task-specific accuracy, a holistic approach considers system-level performance, including scalability, efficiency, and robustness, ensuring that models are not only effective but also practical for real-world deployment.

MMBench is a specialized benchmarking designed for vision-language tasks, providing a holistic evaluation framework that goes beyond traditional benchmarks focused on isolated metrics or narrow datasets. MMBench implements a robust CircularEval strategy and leverages large language models to transform free-form predictions into predefined options, ensuring precise evaluation results, even for models with limited instruction-following abilities. CircularEval is able to achieve a good trade-off between robustness and cost. It addresses limitations in existing benchmarks by providing over 3,000 multiple-choice questions spanning 20 ability dimensions. Those 20 dimensions are branched from reasoning and perception tasks. 8 dimensions are branched from reasoning, including physical property, function relation, identity reasoning (above belongs to attribute reasoning), future prediction, structuralized image-text understanding (above belongs to logical reasoning), social relation, physical relation, and natural relation (above belongs to relation reasoning). 12 dimensions lie under the perception category, including image topic, image quality, image emotion, image scene, image style (above belongs to coarse perception), spatial relationship, attribute comparison, action recognition (above belongs to cross instance fine-grained perception), attribute recognition, object localization, celebrity recognition, OCR (above belongs

to single-task fine-grained perception). This comprehensive scoreboard provides a more holistic evaluation of the VLM performance and provides more robust results compared with other single-task benchmarks.

MMStar [81] is a meticulously curated multi-modal benchmark designed to evaluate the capabilities of large vision-language models (LVLMs). It addresses critical issues in existing benchmarks, such as visual redundancy and data leakage, by ensuring that its 1,500 evaluation samples demand genuine visual dependency and advanced multi-modal reasoning. MMStar spans six core capabilities and 18 detailed axes, rigorously assessing areas like fine-grained perception, logical reasoning, and scientific analysis. The benchmark includes innovative metrics, such as Multi-modal Gain and Multi-modal Leakage, to accurately measure the effectiveness of multi-modal training while mitigating biases from leaked training data. MMStar provides a robust, balanced, and high-quality evaluation platform, offering researchers a clear lens to assess and improve the performance of LVLMs.

MME [82] is the first comprehensive evaluation benchmark for Multimodal Large Language Models (MLLMs), designed to assess both perception and cognitive abilities across 14 subtasks. The benchmark addresses limitations in existing evaluation methods by using entirely manually designed instruction-answer pairs, which eliminates the risk of data leakage from publicly available datasets. MME features concise and standardized instructions, allowing fair comparisons among MLLMs without reliance on extensive prompt engineering. By evaluating 30 advanced MLLMs, MME identifies significant gaps and potential directions for improvement in tasks such as commonsense reasoning, numerical calculation, and code reasoning. This benchmark sets a new standard for rigorously testing and advancing the capabilities of MLLMs.

MMMU [83] (Massive Multi-discipline Multimodal Understanding and Reasoning) is an extensive benchmark tailored for assessing the expert-level capabilities of large multimodal models (LMMs). It features 11,500 college-level questions sourced from exams, textbooks, and quizzes, spanning six core disciplines—Art Design, Business, Science, Health Medicine, Humanities Social Sciences, and Technology Engineering—covering 30 subjects and 183 subfields. MMMU challenges models with highly heterogeneous data, including diagrams, tables, medical images, and more, demanding deep domain knowledge and reasoning. Unlike existing benchmarks focused on commonsense reasoning, MMMU emphasizes complex perception and deliberate reasoning. Evaluations highlight significant gaps between current model performance and human expertise, positioning MMMU as a vital tool for driving progress toward advanced AI capable of handling domain-specific multimodal tasks.

MM-Vet [84] is a benchmark designed to evaluate the integrated capabilities of Large Multimodal Models across complex tasks requiring combinations of six core vision-language abilities: recognition, OCR, knowledge, language generation, spatial awareness, and math. It includes 16 integrated tasks and employs an innovative LLM-based evaluator to score open-ended model outputs, ensuring thorough assessment across diverse question types and answer styles. MM-Vet provides insights beyond simple performance rankings by analyzing per-capability strengths and weaknesses, facilitating detailed comparisons of different LMM designs and paradigms. This benchmark sets a new standard for evaluating the multi-faceted abilities of LMMs and identifying areas for advancement in multimodal AI.

Many of these cutting-edge benchmarks are used to rank the performance of VLMs in various public leaderboards. This fact also shows the effectiveness of comprehensive benchmarks. We will introduce one of the most famous leaderboards OpenVLM [85], presented by HuggingFace, in the next section.

F. HuggingFace OpenVLM Leaderboard

HuggingFace manages a public real-time leaderboard for open-source VLMs and API models. The current OpenVLM leaderboard [85] covers 169 different VLMs, including the most famous ones GPT-4v, Gemini, QwenVLPlus, etc. The leaderboard integrates 31 different multi-modal benchmarks, and users can customize the selection of benchmarks they want to use in their analysis, generating rankings and average scores. The OpenVLM leaderboard is supported by VLMEvalKit. The default selection of benchmarks for evaluating the average performance score is comprehensive, covering evaluations on general ability, math reasoning, hallucination, etc. The result shown below uses the same selection, included benchmarks are: MMBench-V11 [80], MMStar [81], MMMU-VAL [83], MathVista [73], OCRBench [86], AI2D [87], HallusionBench [78], and MMVet [84]. For the convenience of fitting the table in the paper, only the score on MMBench-V11 and MMStar is displayed in the table, as a reference to the most important metric Avg Score. Scores are normalized to a scale of 0-100, and the Avg Score represents the average score on all the benchmarks selected. A higher score represents better performance. Moreover, the table only keeps the best model within the same model family to make better comparisons between different model families and training schemes. For example, Qwen2-VL-72B and Qwen-VL-Max-0809 take the first and second but both of them belong to the Qwen2-72B model family, so we only keep the score and rank of Qwen2-VL-72B to keep the table clear.

V. Conclusion

The development of Vision-Language Models has witnessed remarkable progress, driven by innovations in architecture, training paradigms, and evaluation strategies, with state-of-the-art models demonstrating capabilities in a wide array

Rank	Method	Avg Score	MMBench V11	MMStar
1	Qwen2-VL-72B	74.8	85.9	68.6
2	Step-1.5V	72.5	82	65.1
3	GPT-4o (0806)	71.5	80.5	64.7
4	Ovis1.6-Gemma2	71.3	82.2	63.5
5	InternVL2-Llama3	71	85.5	67.1
6	Claude3.5-Sonnet	70.6	81.7	65.1
7	TeleMM	69.6	82.7	67.9
8	JT-VL-Chat-V3.0	68.4	82.9	65.9
9	LLaVA-OneVision	68	84.5	65.8
10	bailingMM-mini	67	82.2	61.3

TABLE I
VLM LEADERBOARD

of tasks that combine visual and textual modalities. These advancements highlight the potential of VLMs in areas such as multimodal reasoning, open-world recognition, and cross-disciplinary problem-solving.

The evaluation landscape for VLMs is evolving, with benchmarks such as MM-Vet, MMStar, and MME driving the focus toward comprehensive assessments of integrated abilities and fair comparisons. These benchmarks expose gaps between current model performance and human-level understanding, emphasizing the need for better architectures, more diverse and high-quality training datasets, and improved evaluation strategies.

Challenges remain in the current stage of VLM development, including but not limited to scalability issues due to high computational demands, biases in training data that hinder fairness and generalization, and limited robustness to noisy or adversarial inputs. Future VLM development could focus on refining model robustness, reducing reliance on large-scale resource consumption, and enhancing interpretability to ensure these systems are trustworthy and applicable across critical domains. Addressing these challenges will enable VLMs to transition from promising prototypes to transformative tools in industries. Furthermore, VLMs could be potentially extended to video modality, which will unveil much broader application scenarios.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[4] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019. [Online]. Available: https://arxiv.org/abs/1908.02265

[5] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," 2019. [Online]. Available: https://arxiv.org/abs/1908.07490

[6] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[7] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021. [Online]. Available: https://arxiv.org/abs/2102.05918

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[10] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," 2018. [Online]. Available: https://arxiv.org/abs/1707.07998

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016. [Online]. Available: https://arxiv.org/abs/1506.01497

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018. [Online]. Available: https://arxiv.org/abs/1703.06870

[13] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," 2018. [Online]. Available: https://arxiv.org/abs/1803.08024

[14] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," 2019. [Online]. Available: https://arxiv.org/abs/1908.03557

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[16] R. Girshick, "Fast r-cnn," 2015. [Online]. Available: https://arxiv.org/abs/1504.08083

[17] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," 2020. [Online]. Available: https://arxiv.org/abs/1909.11740

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers distillation through attention," 2021. [Online]. Available: https://arxiv.org/abs/2012.12877

[20] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," 2021. [Online]. Available: https://arxiv.org/abs/2102.03334

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," 2016. [Online]. Available: https://arxiv.org/abs/1610.02055

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1711.07971

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014. [Online]. Available: https://arxiv.org/abs/1311.2524

[25] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" 2022. [Online]. Available: https://arxiv.org/abs/2108.08810

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[27] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," 2021. [Online]. Available: https://arxiv.org/abs/2107.07651

[28] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes,

Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," 2022. [Online]. Available: https://arxiv.org/abs/2204.02311

[29] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "Palm-e: An embodied multimodal language model," 2023. [Online]. Available: https://arxiv.org/abs/2303.03378

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2015. [Online]. Available: https://arxiv.org/abs/1411.4555

[31] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," 2015. [Online]. Available: https://arxiv.org/abs/1412.2306

[32] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," 2015. [Online]. Available: https://arxiv.org/abs/1412.6632

[33] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," 2016. [Online]. Available: https://arxiv.org/abs/1411.4389

[34] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014. [Online]. Available: https://arxiv.org/abs/1406.1078

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2016. [Online]. Available: https://arxiv.org/abs/1502.03044

[36] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," 2019. [Online]. Available: https://arxiv.org/abs/1906.10770

[37] M. Stefanini, M. Cornia, L. Baraldi, and R. Cucchiara, "A novel attention-based aggregation function to combine vision and language," 2020. [Online]. Available: https://arxiv.org/abs/2004.13073

[38] E. Bugliarello, R. Cotterell, N. Okazaki, and D. Elliott, "Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts," 2021. [Online]. Available: https://arxiv.org/abs/2011.15124

[39] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[40] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[41] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman *et al.*, "An introduction to vision-language modeling," *arXiv preprint arXiv:2405.17247*, 2024.

[42] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang *et al.*, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006.

[43] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[44] T. Mikolov, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, vol. 3781, 2013.

[45] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[47] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[48] S. Jang, J. Yun, J. Kwon, E. Lee, and Y. Kim, "Dial: Dense image-text alignment for weakly supervised semantic segmentation," *arXiv preprint arXiv:2409.15801*, 2024.

[49] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[50] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[51] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[52] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 638–15 650.

[53] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *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*. Springer, 2020, pp. 121–137.

[54] OpenAI, "Dall·e: Creating images from text," https://openai.com/index/dall-e/, accessed: Dec. 10, 2024.

[55] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[56] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[57] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[58] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[59] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[60] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[61] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[62] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[63] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," 2015. [Online]. Available: https://arxiv.org/abs/1504.00325

[64] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[65] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, and L. Specia, Eds. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 376–380. [Online]. Available: https://aclanthology.org/W14-3348

[66] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for

Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[67] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," 2015. [Online]. Available: https://arxiv.org/abs/1411.5726

[68] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," 2017. [Online]. Available: https://arxiv.org/abs/1612.00837

[69] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," 2016. [Online]. Available: https://arxiv.org/abs/1612.06890

[70] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," 2019. [Online]. Available: https://arxiv.org/abs/1902.09506

[71] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," 2018. [Online]. Available: https://arxiv.org/abs/1802.08218

[72] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," *CoRR*, vol. abs/1906.00067, 2019. [Online]. Available: http://arxiv.org/abs/1906.00067

[73] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," 2024. [Online]. Available: https://arxiv.org/abs/2310.02255

[74] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, P. Gao, and H. Li, "Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?" 2024. [Online]. Available: https://arxiv.org/abs/2403.14624

[75] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," 2019. [Online]. Available: https://arxiv.org/abs/1809.02156

[76] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.10355

[77] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multi-modal models via robust instruction tuning," 2024. [Online]. Available: https://arxiv.org/abs/2306.14565

[78] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou, "Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," 2024. [Online]. Available: https://arxiv.org/abs/2310.14566

[79] X. Wu, T. Guan, D. Li, S. Huang, X. Liu, X. Wang, R. Xian, A. Shrivastava, F. Huang, J. L. Boyd-Graber, T. Zhou, and D. Manocha, "Autohallusion: Automatic generation of hallucination benchmarks for vision-language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.10900

[80] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "Mmbench: Is your multi-modal model an all-around player?" 2024. [Online]. Available: https://arxiv.org/abs/2307.06281

[81] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, and F. Zhao, "Are we on the right way for evaluating large vision-language models?" 2024. [Online]. Available: https://arxiv.org/abs/2403.20330

[82] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji, "Mme: A comprehensive evaluation benchmark for multimodal large language models," 2024. [Online]. Available: https://arxiv.org/abs/2306.13394

[83] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," 2024. [Online]. Available: https://arxiv.org/abs/2311.16502

[84] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," 2024. [Online]. Available: https://arxiv.org/abs/2308.02490

[85] O. Contributors, "Opencompass: A universal evaluation platform for foundation models," https://github.com/open-compass/opencompass, 2023.

[86] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X. Yin, C. lin Liu, L. Jin, and X. Bai, "Ocrbench: On the hidden mystery of ocr in large multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2305.07895

[87] T. Hiippala, M. Alikhani, J. Haverinen, T. Kalliokoski, E. Logacheva, S. Orekhova, A. Tuomainen, M. Stone, and J. A. Bateman, "Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams," *Language Resources and Evaluation*, vol. 55, no. 3, p. 661–688, Dec. 2020. [Online]. Available: http://dx.doi.org/10.1007/s10579-020-09517-1