AIRBUS

Multimodal Representation Learning and Large Vision-Language Models:

A Comprehensive Bibliographic Review

Bilal Azdad

June 15, 2025

Contents

Abstract

The synthesis of information from disparate sources, such as images and text, represents a fundamental challenge and a pivotal area of research in modern artificial intelligence. This study provides an in-depth bibliographic review of multimodal representation learning and the architectures of the large vision-language models (VLMs) that have come to dominate the field. We begin by examining contrastive learning as a foundational methodology for aligning different modalities within a common latent space. Subsequently, we explore the landscape of generative models, covering both text-to-image synthesis and image-to-text generation, with a detailed case study on the architectural evolution of Visual Question Answering (VQA) systems. We then survey the emergence of multimodal large language models (LMMs), analyzing key architectural strategies for fusing visual data with pre-trained language models. Finally, we address the critical and practical challenge of efficiently adapting these billion-parameter models, with a focus on parameter-efficient fine-tuning (PEFT) techniques like Low-Rank Adaptation (LoRA).

1 Introduction

The fields of computer vision (CV) and natural language processing (NLP) have historically progressed along largely independent research trajectories. However, achieving a more generalized form of artificial intelligence necessitates the integration of these modalities to process and reason about information in a manner that mirrors human perception. Vision-Language Models (VLMs) have emerged as the central paradigm for addressing this challenge [1]. A quintessential task motivating this research is Visual Question Answering (VQA), which requires a system to comprehend an image and answer a natural language question about its content, thereby demanding a deep synthesis of visual and linguistic understanding [2].

The core technical obstacle is the development of latent representations that can effectively bridge modalities. This involves mapping high-dimensional inputs like images and text into a shared embedding space where semantic relationships can be evaluated directly [3]. Progress in this domain has been accelerated by key innovations in deep learning, including **contrastive representation learning** [4], the development of **large-scale generative models** [5], and the extension of **transformer-based large language models (LLMs)** to handle multimodal inputs.

While the scale of these models has unlocked unprecedented capabilities, it has also introduced significant practical challenges related to training and adaptation. In response, the research community has developed a suite of parameter-efficient fine-tuning (PEFT) techniques designed to adapt massive models to new tasks with substantially lower computational costs [6]. This review will provide a structured exploration of these topics, examining the foundational methods, model architectures, and training strategies that define the current state of vision-language research.

2 Foundational Methods: Contrastive Representation Learning

A primary goal in multimodal learning is the creation of a unified embedding space where representations from different modalities are directly comparable. Contrastive learning has proven to be a highly effective and scalable framework for achieving this alignment [4]. The central principle is to train encoders that map inputs into this shared space under a simple constraint: the embeddings of semantically related (positive) pairs should be closer together than those of unrelated (negative) pairs. This objective incentivizes the model to learn high-level, modality-invariant semantic features.

2.1 The InfoNCE Objective

Many contrastive learning methods utilize an objective function derived from Noise-Contrastive Estimation (NCE), commonly known as the InfoNCE loss. Given an anchor sample x (e.g., an image), a corresponding positive sample x^+ (e.g., its caption), and a set of M negative samples $\{x_i^-\}$ (e.g., other captions in the training batch), the loss is formulated as:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(f(x), f(x^+))/\tau)}{\exp(\text{sim}(f(x), f(x^+))/\tau) + \sum_{i=1}^{M} \exp(\text{sim}(f(x), f(x_i^-))/\tau)} \tag{1}$$

In this formulation, $f(\cdot)$ represents the encoder network that produces a normalized embedding vector, $\text{sim}(u, v)$ is a similarity metric (typically cosine similarity), and τ is a temperature hyperparameter that modulates the separation of classes [4]. Minimizing this loss encourages the similarity of the positive pair to be significantly higher than that of any negative pair. The use of large batch sizes is a common strategy to ensure a diverse set of challenging negatives.

2.2 Case Study at Scale: The CLIP Model

The CLIP (Contrastive Language-Image Pre-training) model from Radford et al. (2021) represents a landmark application of this paradigm at an unprecedented scale [7]. CLIP consists of two separate encoders—a Vision Transformer (ViT) for images and a text Transformer—jointly trained on a dataset of 400 million image-caption pairs sourced from the web [7].

During training, an $N \times N$ similarity matrix is computed for a batch of N image-text pairs. The model is optimized via a symmetric cross-entropy loss across this matrix, effectively training each image encoder to identify its corresponding text embedding from N candidates, and vice-versa [7]. This dual-objective process results in a highly robust and semantically rich joint embedding space.

The most significant outcome of CLIP's training is its remarkable **zero-shot classification** capability. As depicted in Figure 1, a classifier can be constructed dynamically by creating descriptive prompts for target categories (e.g., "a photo of a {object}"). The model classifies a given image by finding the text prompt whose embedding has the highest cosine similarity to the image's embedding. This enables CLIP to perform classification on arbitrary tasks without any task-specific fine-tuning [7]. While highly effective, CLIP exhibits limitations in tasks requiring very fine-grained discrimination or complex spatial

reasoning, indicating that some granular visual detail may not be perfectly aligned with language [7].

The power of CLIP stems from its use of natural language as a supervision signal. Instead of being trained on one-hot labels, the model learns from descriptive captions, allowing it to generalize to concepts involving style, context, and abstract themes far beyond simple object identity [7]. This has led to further research extending the paradigm. ALIGN, for instance, scaled the approach to an even larger, albeit noisier, dataset of over one billion pairs, while ImageBind demonstrated the potential for a single unified embedding space across seven modalities, including audio and depth [3].

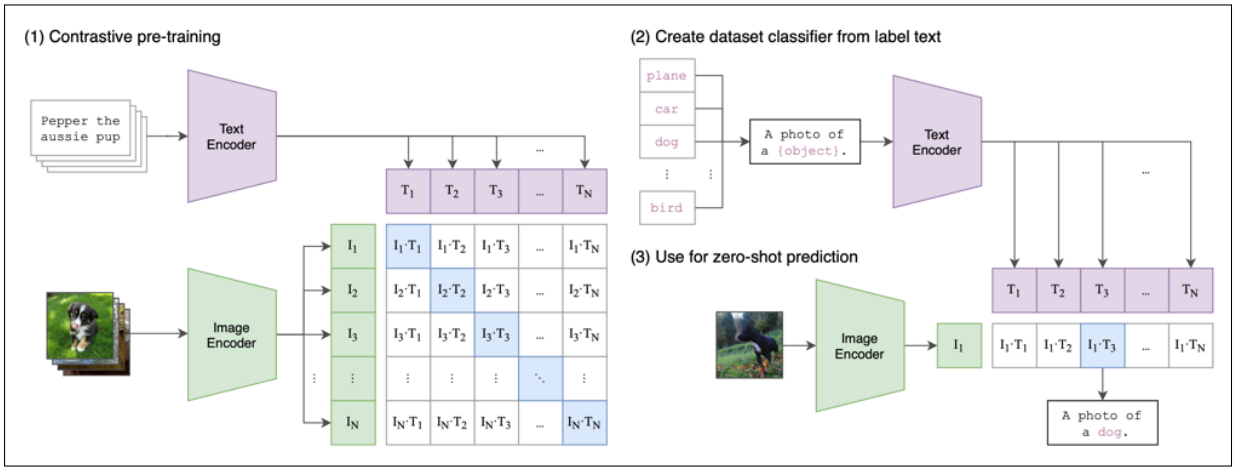

Figure 1: A diagram of the CLIP model's workflow. (1) During contrastive pre-training, image and text encoders are optimized to maximize the similarity of true pairs (diagonal) in an $N \times N$ comparison matrix. (2-3) For zero-shot inference, a target image is classified by comparing its embedding to the embeddings of dynamically created text prompts, selecting the most similar pairing. (Figure adapted from Radford et al., 2021).

2.3 Beyond Simple Pairs: Structured and End-to-End Alignment

While CLIP's approach was powerful, follow-up research sought to achieve a more fine-grained alignment. Early transformer-based models like UNITER and Oscar fused visual and textual features but often relied on pre-extracted region features from object detectors like Faster R-CNN [1]. This two-stage process, where objects are detected first and then aligned with text, could constrain the model's expressiveness by missing finer details not captured in the initial object proposals. Oscar, for its part, attempted to improve this by using detected object tags as "anchor points" to better ground the language in the visual content [1].

The field has increasingly moved towards end-to-end alignment using vision transformers, where the model learns to align text phrases directly with image patches. This is often achieved through cross-attention mechanisms, where words in the text can directly "attend to" different regions of the image, enabling a much more granular and flexible alignment than what is possible with global embedding similarity alone.

3 Generative Modeling in Vision-Language Systems

While contrastive methods excel at learning unified representations, a separate and highly active research direction focuses on generative modeling: creating novel content from multimodal inputs. This area can be broadly divided into text-to-image and image-to-text generation.

3.1 Text-to-Image Synthesis

This subfield has undergone rapid evolution, moving from early GAN-based methods (e.g., StackGAN) to large-scale Transformer and diffusion models [5].

- **Autoregressive Models:** OpenAI's DALL·E was a pioneering 12-billion parameter model that framed text-to-image generation as a language modeling task. It tokenized both text and images into a single sequence using a discrete Vector Quantised-Variational AutoEncoder (VQ-VAE) [8] and trained a large Transformer to predict the next token autoregressively [9]. This unified stream demonstrated strong compositional generalization and enabled creative capabilities like in-painting [9].

- **Diffusion Models:** More recent state-of-the-art models, such as DALL·E 2 and Stable Diffusion, are based on the diffusion paradigm. These models learn to reverse a gradual noising process, generating an image by iteratively de-noising a random seed, conditioned on a text embedding (often from a pre-trained CLIP model) [5,10]. DALL·E 2, for example, uses a two-stage approach: first generating a CLIP image embedding from the text, then using a diffusion decoder to produce a high-resolution image matching that embedding. Diffusion models are known for their high-fidelity output and training stability, and have become the dominant architecture for this task. This frontier is now rapidly expanding from static images to dynamic content, with models like Google's Veo 3 generating high-definition video and synchronized audio from a single text prompt. [9].

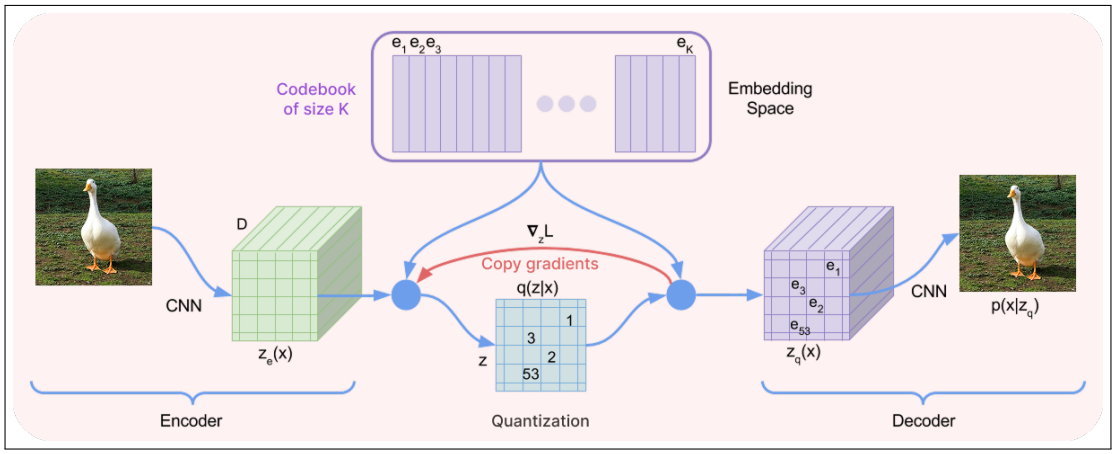

Figure 2: Conceptual structure of a VQ-VAE. The encoder output $z_e(x)$ is mapped to the nearest entry e_k in a learned codebook. The gradient from the decoder is copied directly to the encoder output via a straight-through estimator, enabling end-to-end training. [8]

3.2 Image-to-Text Generation and Visual Reasoning

Generating textual output from visual input encompasses tasks from image captioning to VQA. Early captioning models (e.g., Show-and-Tell) utilized a CNN-RNN encoder-decoder structure [1]. The introduction of attention mechanisms in models like Show-Attend-and-Tell was a key improvement, allowing the decoder to focus on relevant image regions during text generation [1].

VQA presents a more complex challenge that serves as a benchmark for multimodal reasoning. The evolution of VQA architectures reveals a clear progression:

1. **Early Fusion Models (c. 2015-2017):** These models employed simple fusion techniques, such as element-wise multiplication, to combine global image and question features before classification [1].

2. **Attention-Based Models:** Subsequent models, like Stacked Attention Networks, integrated attention mechanisms, enabling the model to dynamically weight different image regions based on the content of the question, thereby focusing on relevant visual evidence [1].

3. **Object-Centric Approaches:** The "bottom-up and top-down attention" model by Anderson et al. became highly influential. It used a pre-trained object detector (Faster R-CNN) to propose salient regions (bottom-up), which were then attended to by a language-guided mechanism (top-down) [1]. This object-centric view dramatically improved performance.

4. **Unified Transformer Models:** More recent models (e.g., UNITER, LXMERT, ViLBERT) leverage Transformers to jointly reason over a sequence of image-region and text-word embeddings. Pre-training on large-scale, general-purpose datasets with objectives like masked language/region modeling allows these models to transfer rich world knowledge to the VQA task, overcoming some of the language biases and reasoning limitations of earlier approaches [1].

Figure 3: An illustrative VQA task. Answering questions about the content of an image requires a model to perform a sequence of sub-tasks, including object recognition, attribute identification, spatial reasoning, and commonsense inference. (Figure adapted from Viso.ai, 2024).

4 The Emergence of Large Multimodal Models (LMMs)

A recent paradigm shift involves integrating vision capabilities directly into pre-trained LLMs, creating what are now known as Large Multimodal Models (LMMs). GPT o3 or Gemini 2.5 pro, which accept both image and text inputs, are a prominent example of models that exhibit human-level performance on various academic benchmarks by leveraging this fusion [11]. This approach aims to synergize the powerful reasoning and world knowledge inherent in LLMs with perceptual capabilities.

4.1 Architectural Strategies for Modality Fusion

The central design challenge for LMMs is the interface between the vision encoder and the LLM. Several effective strategies have emerged:

- **Prefix-Tuning and Prompting (Flamingo):** This method keeps the LLM entirely frozen. A small, trainable network, the Perceiver Resampler in Flamingo's case, processes the visual features and produces a fixed number of "soft prompt" embeddings. These visual embeddings are then prepended or interleaved with the textual input to guide the LLM's generation without modifying its weights [3].
- **Bridge Networks (BLIP-2):** This approach introduces a lightweight "bridge" network to connect a frozen image encoder and a frozen LLM. The Q-Former in BLIP-2 is a trainable module designed to extract a small, fixed number of the most salient visual features for the LLM [12]. This architecture is highly parameter-efficient; BLIP-2 was shown to outperform the much larger Flamingo model on zero-shot VQA while using 54 times fewer trainable parameters, highlighting the value of this approach [12].
- **Instruction Tuning (LLaVA):** To enhance conversational ability and alignment, LMMs can be instruction-tuned, a technique analogous to that used for models like ChatGPT. Models like LLaVA and InstructBLIP are fine-tuned on datasets of image-based, instruction-following conversations, teaching them to function as helpful visual assistants rather than simple VQA systems [3].

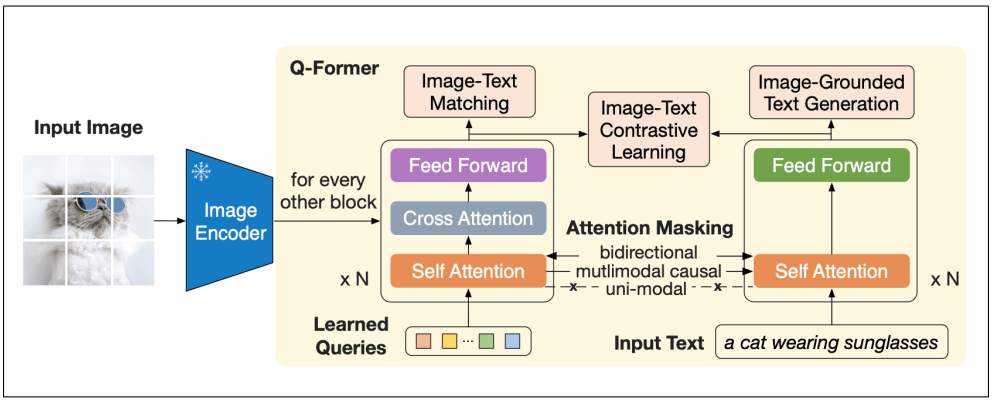

Figure 4: BLIP-2. The image transformer extracts a fixed number of output features from the image encoder, independent of input image resolution, and receives learnable query embeddings as input. The queries can additionally interact with the text through the same self-attention layers.(Figure adapted from Huggingface.com, 2024).

5 Efficiency and Scalability: Parameter-Efficient Fine-Tuning

The immense scale of today's foundation models—often exceeding 10^{11} parameters and requiring super-computers to train end-to-end— makes full fine-tuning out of reach for most research labs and companies [11]. **Parameter-Efficient Fine-Tuning (PEFT)** tackles this bottleneck by updating *only a sliver* of model weights (or auxiliary parameters) while keeping the frozen backbone intact.

5.1 The LoRA Method and Its Variants

Low-Rank Adaptation (LoRA). LoRA [6] leaves the original weight matrix $W \in \mathbb{R}^{d\times d}$ untouched and learns a low-rank update $\Delta W = B\,A$ with $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$, $r \ll d$:

$$\mathbf{h} \;=\; (W + \Delta W)\,\mathbf{x} \equiv (W + B\,A)\,\mathbf{x}, \qquad \operatorname{rank}(\Delta W) = r. \tag{2}$$

Only A and B are trained, cutting the update budget from d^2 to $2dr$. On GPT-3 (175B) with $r = 4$ this is ~18M extra weights—0.01% of the model—yet matches full fine-tuning on many benchmark tasks [6]. As noted by IBM researchers, LoRA makes fine-tuning large models substantially more accessible [13].

Other PEFT Families. Other methods offer different trade-offs. **IA**^{3} [14] introduces even fewer parameters by learning scaling vectors that multiply the outputs of feed-forward and attention layers. **Adapter modules** [15] insert small, trainable bottleneck MLPs between a model's layers. **Prompt and prefix tuning** [16] learn continuous task-specific vectors that are prepended to the input sequence, leaving the base model entirely untouched. Finally, **BitFit** updates only the bias terms of the network, proving surprisingly effective for its simplicity [17].

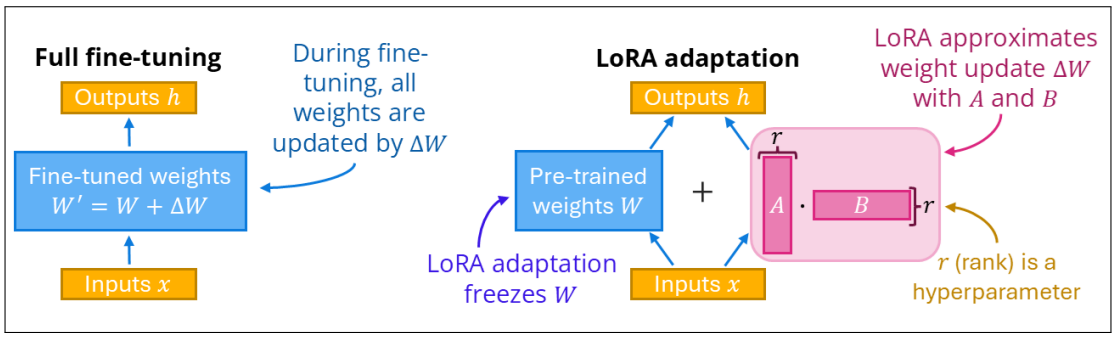

Figure 5: Low-Rank Adaptation (LoRA) injected into a single Transformer weight matrix. The pre-trained weights W remain frozen while the low-rank matrices A and B are trained. [18]

5.2 Quantisation and QLoRA

To further reduce the memory footprint, PEFT methods can be combined with quantization. **QLoRA** couples 4-bit NormalFloat quantization with low-rank updates, storing the large, frozen base weights in 4-bit precision while training the much smaller LoRA adapter matrices in a higher precision (e.g., 16-bit) [19]. This technique can yield memory reductions of up to 16x, enabling tasks like fine-tuning a 65-billion parameter model on a single 24GB consumer GPU.

5.3 Implications for Multimodal Research

PEFT techniques are pivotal for the practical development of LMMs. A common recipe involves:

- **Freezing the vision stack** (e.g., a pre-trained ViT) and the **language model backbone**.
- **Applying LoRA or adapters** only to the cross-modal "bridge" network (e.g., BLIP-2's Q-Former) or to selected layers of the language decoder.
- **Combining with 4/8-bit quantisation** to ensure that even massive checkpoints can fit onto commodity hardware for training and inference.

This strategy has underpinned the rapid development and release of powerful open-source community models like MiniGPT-4, Qwen-VL, and InternVL, dramatically accelerating the research cycle.

6 Conclusion and Future Directions

The field of multimodal machine learning has made significant strides, moving from siloed models to deeply integrated systems capable of sophisticated cross-modal reasoning and generation. The progress driven by contrastive learning, generative architectures, and efficient fine-tuning techniques is now paving the way for applications far beyond academic benchmarks. For an enterprise such as Airbus, the abstract challenges of multimodal research translate into tangible operational goals, moving beyond isolated capabilities toward a unified, context-aware system.

The Industrial Digital Twin. The future lies in creating a holistic model of an asset, such as an aircraft. An advanced VLM could fuse real-time video feeds from a maintenance check with 3D CAD schematics and historical repair logs. This would empower an engineer to query the system in natural language—"Show all documented stress fractures on this fuselage section and cross-reference with the original manufacturing specifications"—and receive a synthesized, multimodal response.

Expert Assistance and Embodied AI. This same model could act as an expert assistant, guiding a technician through a complex repair by visually identifying components and displaying the relevant procedural steps. As a natural extension, this intelligence could be integrated into embodied agents, enabling robots to perform intricate assembly or inspection tasks with a degree of autonomy previously unattainable.

The Path to Industrial Adoption. However, transitioning these models from research prototypes to certified industrial tools presents the next great challenge for the field. The focus must shift from pure performance metrics to demonstrable reliability, transparency, and security. In an environment where operational safety is non-negotiable, the key research questions become: How can we guarantee the factuality of a model's output? How can we audit its reasoning process? How do we ensure its robustness against unforeseen edge cases? Future work will therefore be defined not only by the pursuit of more capable models, but by the rigorous engineering required to make them **certifiably safe and trustworthy** partners in the world's most demanding applications.

References

[1] A. Pandey et al., 2025. *The Quest for Visual Understanding: A Journey Through the Evolution of Visual Question Answering*. arXiv:2501.07109v1.

[2] Viso.ai Blog, 2024. *Understanding Visual Question Answering (VQA) in 2025*. Published online at viso.ai.

[3] C. Huyen, 2023. *Multimodality and Large Multimodal Models (LMMs)*. Blog post, published online at huyenchip.com.

[4] L. Weng, 2021. *Contrastive Representation Learning*. Blog post, Lil'Log.

[5] C. Zhang et al., 2023. *Text-to-image Diffusion Models in Generative AI: A Survey*. arXiv:2303.07909v3.

[6] E. J. Hu et al., 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. Published in ICLR, 2022.

[7] A. Radford et al., 2021. *Learning Transferable Visual Models From Natural Language Supervision*. Published in ICML, 2021.

[8] A. van den Oord, O. Vinyals, & K. Kavukcuoglu, 2017. *Neural Discrete Representation Learning*. Published in NeurIPS, 2017.

[9] OpenAI, 2021. *DALL·E: Creating Images from Text*. OpenAI Blog.

[10] Myriad, 2023. *DALL-E 2 explained*. Blog post.

[11] OpenAI, 2023. *GPT-4 Technical Report*. arXiv:2303.08774.

[12] J. Li et al., 2023. *BLIP-2: Bootstrapping Language-Image Pre-training with a Frozen Image Encoder and a Frozen Large Language Model*. Published in ICML, 2023.

[13] IBM Research, 2023. *What is LoRA (Low-Rank Adaption)?*. IBM Research Blog.

[14] H. Liu et al., 2022. *Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning*. Published in NeurIPS, 2022. (Paper associated with IA3).

[15] N. Houlsby et al., 2019. *Parameter-Efficient Transfer Learning for NLP*. Published in ICML, 2019.

[16] P. Liu et al., 2023. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. ACM Computing Surveys, 55(9).

[17] E. Ben Zaken, S. Ravfogel, & Y. Goldberg, 2022. *BitFit: Simple Parameter-Efficient Fine-Tuning for Transformers*. Published in ACL, 2022.

[18] D. Herrmannova, 2023. *Low-Rank Adaptation (LoRA)*. One-Minute NLP substack post.

[19] T. Dettmers et al., 2023. *QLoRA: Efficient Finetuning of Quantized LLMs*. Published in NeurIPS, 2023.

[20] OpenAI, 2023. *GPT-4V(ision) System Card*. PDF document, published online.