An In-Depth Survey of Multimodal Foundation Models and Their Challenges

Haoran Yijun[1], Shufen Zhihao[1]

[1]Department of Computer Science and Technology, Fudan University, China

Abstract

Multimodal foundation models have emerged as a transformative paradigm in artificial intelligence, enabling the integration and joint understanding of heterogeneous data modalities such as vision, language, audio, and beyond. These models leverage large-scale pretraining on massive, diverse multimodal datasets to learn rich, transferable representations that underpin a wide spectrum of downstream tasks, including retrieval, generation, classification, and reasoning. This survey provides a comprehensive overview of the current landscape of multimodal foundation models, tracing key trends in architecture design, cross-modal alignment, fusion techniques, and training methodologies. We discuss prominent evaluation benchmarks and metrics that assess performance, robustness, and fairness across multimodal tasks. Furthermore, we analyze critical challenges such as modality heterogeneity, scalability, interpretability, and ethical considerations that remain barriers to widespread adoption. Finally, we highlight emerging opportunities and future directions, including unified multimodal architectures, continual learning, and responsible AI practices. Our goal is to offer a unified and in-depth resource that elucidates the theoretical foundations, practical implementations, and societal implications of multimodal foundation models, thereby guiding future research and development in this rapidly evolving field.

Keywords: Multimodal Foundation Models, Large Language Models, Large Vision Models, Prompt Tuning, Adapter Tuning, LoRA, Low-Rank Adaptation, BitFit, Sparse Fine-Tuning, Transfer Learning, Multimodal Models, Efficient Adaptation, Model Compression, Continual Learning.

1 Introduction

The recent advent of foundation models has significantly reshaped the landscape of artificial intelligence (AI), especially through their capacity to generalize across a

wide range of downstream tasks via pretraining on massive datasets [1]. While early successes of foundation models have been concentrated in unimodal domains—most notably large language models (LLMs) such as GPT, PaLM, and LLaMA for text, or vision transformers (ViTs) like BEiT and MAE for images—the focus has rapidly shifted towards building models capable of processing and integrating multiple modalities, giving rise to the emergent field of multimodal foundation models [2]. These models, such as CLIP, Flamingo, Gato, and GPT-4, operate across text, vision, audio, and even action or sensory modalities, offering a unified computational interface to heterogeneous inputs [3]. Multimodal foundation models promise to unlock a richer form of artificial general intelligence (AGI) by simulating human-like perception and reasoning capabilities that are inherently grounded in multi-sensory experience [4]. However, developing such models introduces a multitude of challenges not present in unimodal settings, such as learning aligned cross-modal representations, handling heterogeneous data structures and temporal scales, ensuring modality-aware attention and fusion, and mitigating imbalances in data availability and quality across modalities. These complexities demand the development of novel architectures, training objectives, alignment strategies, and evaluation metrics to ensure generalization, interpretability, and robustness in multimodal contexts. At the heart of this movement is the convergence of scalable model architectures, self-supervised learning paradigms, and ever-expanding multimodal datasets [5]. Transformer-based architectures, in particular, have served as the backbone for most recent advances due to their capacity to scale, handle long-range dependencies, and be adapted flexibly to various modalities either through shared or modality-specific tokenization [6]. Self-supervised learning techniques such as contrastive learning, masked modeling, and alignment losses (e.g., CLIP's InfoNCE) have become crucial to learning from weakly labeled or entirely unlabeled multimodal corpora. At the same time, the availability of large-scale web-scraped data—such as image-text pairs from the internet, video-audio transcriptions, and interactive robot-environment logs—has catalyzed the training of general-purpose models capable of zero-shot and few-shot reasoning. However, the heterogeneity, noise, and bias inherent in such data raise critical concerns regarding fairness, safety, and the true extent of generalization. Unlike their unimodal counterparts, multimodal foundation models must reconcile semantic representations across vastly different modalities, for example mapping visual scenes to abstract textual descriptions, or interpreting natural language in the context of spatial or temporal sensory data. This adds an additional layer of complexity to architectural design and pretraining objectives, necessitating careful calibration of cross-modal representations to avoid misalignment or modality collapse [7]. In this survey, we aim to provide a comprehensive and structured overview of the rapidly evolving field of multimodal foundation models, encompassing their ar-

chitectural underpinnings, pretraining strategies, alignment techniques, emergent capabilities, and the many open challenges that remain [8]. We categorize the vast design space along several axes: model architecture (encoder-decoder, dual-stream, unified transformer), modality configuration (vision-language, audio-text, video-language, etc.), training paradigms (contrastive, generative, hybrid), and alignment techniques (cross-attention, fusion layers, shared embeddings) [9]. Furthermore, we trace the historical evolution of multimodal learning leading up to the foundation model era, highlighting how earlier systems such as image captioning and VQA evolved into today's unified models. We also examine critical issues such as evaluation benchmarking, generalization across tasks and domains, robustness to modality corruption, compositional reasoning, data efficiency, and ethical risks including bias propagation and misuse. By synthesizing existing methods, emerging trends, and ongoing debates, we intend to provide researchers, practitioners, and policymakers with a unified lens through which to understand the state and trajectory of multimodal foundation models [10]. Ultimately, we envision that continued progress in this field will not only enhance the scope of intelligent systems but also push forward fundamental questions about representation, reasoning, and human-machine interaction in a world of ever-increasing sensory and semantic complexity [11].

2 Architectural Taxonomy of Multimodal Foundation Models

The architectural design of multimodal foundation models is central to their ability to generalize across heterogeneous modalities and tasks [12]. At a high level, most models fall into one of three overarching architectural paradigms: **encoder-only**, **decoder-only**, and **encoder-decoder** architectures [13]. Let $\mathcal{X}^{(v)}$, $\mathcal{X}^{(t)}$, and $\mathcal{X}^{(a)}$ represent input modalities such as vision, text, and audio, respectively [14]. A model learns a joint representation space $\mathcal{Z}$ such that each modality-specific encoder $E^{(m)} : \mathcal{X}^{(m)} \to \mathcal{Z}$ maps raw inputs into a shared latent space, while a decoder $D : \mathcal{Z} \to \mathcal{Y}$ generates outputs in the target space $\mathcal{Y}$, which may be a language token sequence, image, or task-specific output (e.g., bounding boxes, actions) [15]. In encoder-only designs such as CLIP or ALIGN, multiple modality-specific encoders are trained to project data into a common embedding space using contrastive losses like $\mathcal{L}_{\text{InfoNCE}}$ [16]. Formally, given positive pairs $(x_i^{(v)}, x_i^{(t)})$ and negatives $(x_i^{(v)}, x_j^{(t)})$ with $i \neq j$, the model minimizes a loss of the form:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i^{(v)}, z_i^{(t)})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i^{(v)}, z_j^{(t)})/\tau)}$$

where $z_i^{(m)} = E^{(m)}(x_i^{(m)})$ is the modality embedding and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. In contrast, decoder-only models like Flamingo or GPT-4 with multimodal adapters rely on a single transformer that autoregressively generates outputs conditioned on cross-modal context, often relying on a learned projection of non-text modalities into the token space [17]. These models are optimized via next-token prediction with a loss $\mathcal{L}_{\text{LM}} = -\sum_t \log p(y_t | y_{<t}, \mathcal{X})$, where $\mathcal{X}$ includes projected multimodal embeddings prepended or interleaved with text tokens. Lastly, encoder-decoder models such as OFA, GIT, and PaLI introduce a modality-agnostic encoder stack followed by a generative or discriminative decoder, unifying inputs across vision, text, and speech into a dense latent representation that facilitates structured outputs [18]. These architectures often support more flexible task formats, especially when trained with multi-task objectives or prompted via instruction tuning [19].

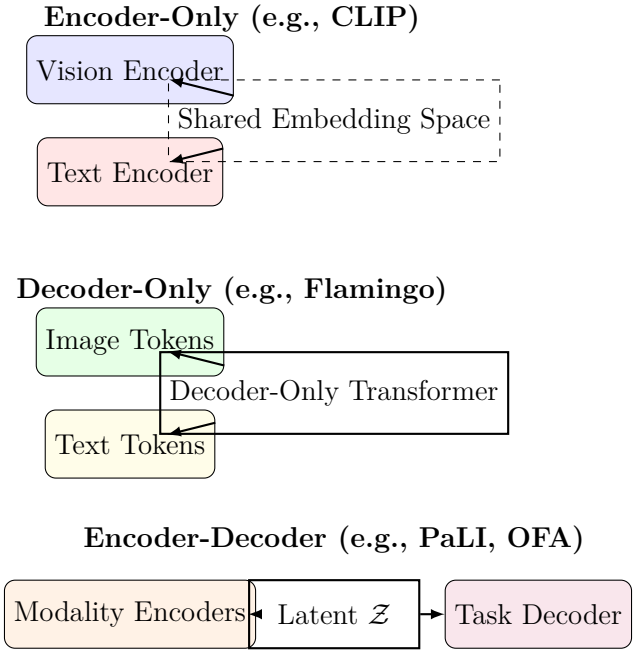

Figure 1: Taxonomy of Multimodal Foundation Model Architectures.

In all cases, a central challenge is the design of modality-specific embeddings and their alignment in a shared semantic space [20]. Whether via late fusion (e.g., projecting final representations into a joint space), early fusion (e.g., input-level concatenation), or intermediate fusion through cross-attention layers, the fidelity of cross-modal integration critically affects the model's generalization and transferability [21]. Importantly, attention mechanisms $\text{Attn}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d})V$ can be modified to include modality-specific gating or learned mixture-of-experts

layers that adaptively weight modal contributions. This flexibility enables both modality-invariant and modality-aware computations depending on the task demands [22]. As we will explore in the next sections, each architectural choice reflects a trade-off among scalability, expressiveness, task alignment, and training complexity.

3 Training Paradigms and Objectives for Multimodal Foundation Models

The effectiveness of multimodal foundation models critically depends not only on their architectural design but also on the choice of training paradigms and learning objectives that enable the model to discover meaningful cross-modal relationships and generalizable representations [23]. Unlike unimodal models, which often rely on a single objective such as language modeling or image classification, multimodal models must contend with heterogeneous data formats, missing modalities, and varying levels of supervision [24]. As a result, contemporary approaches employ a diverse set of training paradigms ranging from self-supervised contrastive learning, masked prediction, and generative modeling, to hybrid multi-task and instruction-tuning strategies. Each paradigm aims to address different facets of the multimodal learning problem: for instance, contrastive learning enforces alignment between paired modalities to learn a shared semantic space, while generative objectives encourage modeling of conditional distributions across modalities, thus supporting complex cross-modal generation and reasoning [25]. Formally, given paired data points $\{(x_i^{(v)}, x_i^{(t)})\}_{i=1}^N$ sampled from a joint distribution $p(x^{(v)}, x^{(t)})$, where $x_i^{(v)} \in \mathcal{X}^{(v)}$ is a visual input and $x_i^{(t)} \in \mathcal{X}^{(t)}$ a corresponding text input, contrastive training methods typically minimize a symmetric InfoNCE loss defined as:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(z_i^{(v)}, z_i^{(t)})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i^{(v)}, z_j^{(t)})/\tau)} + \log\frac{\exp(\text{sim}(z_i^{(t)}, z_i^{(v)})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i^{(t)}, z_j^{(v)})/\tau)}\right]$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, $z_i^{(m)}$ are modality-specific embeddings, and τ is a temperature hyperparameter [26]. This loss encourages the model to bring representations of matching pairs closer while pushing apart mismatched pairs, thus facilitating alignment in the joint embedding space [27]. However, purely contrastive methods may struggle to capture the full generative and semantic richness of cross-modal relationships, especially when the alignment between modalities is noisy or incomplete [28]. To complement contrastive approaches, generative modeling paradigms such as masked language modeling (MLM) and masked visual modeling (MVM) have been integrated into multimodal training objectives

[29]. Masked modeling forces the model to reconstruct missing components of input sequences, encouraging deeper understanding of intra- and inter-modal dependencies [30]. For example, given a text input sequence $X = (x_1, \ldots, x_T)$ with tokens randomly masked at positions $\mathcal{M} \subseteq \{1, \ldots, T\}$, the MLM objective is to maximize the log-likelihood

$$\mathcal{L}_{\text{MLM}} = -\sum_{t \in \mathcal{M}} \log p_\theta(x_t | X_{\setminus \mathcal{M}}, Z^{(v)})$$

where $X_{\setminus \mathcal{M}}$ denotes the unmasked tokens and $Z^{(v)}$ is the visual encoding that conditions text prediction [31]. Similarly, MVM extends this idea to visual tokens or patches, requiring the model to infer masked image regions conditioned on textual context, thus reinforcing cross-modal reasoning. Recent work has also explored unified masked modeling objectives that jointly mask and predict tokens across modalities, promoting synergy between language and vision understanding. Hybrid training strategies combine contrastive, generative, and discriminative losses in multi-task or multi-stage frameworks to leverage the strengths of each approach. For example, a model may be pretrained with contrastive alignment on large-scale noisy image-text pairs, followed by fine-tuning with generative captioning and question-answering objectives that require multi-hop reasoning across modalities [32]. Moreover, instruction tuning with multimodal prompts has emerged as a powerful paradigm for aligning foundation models with human intent, enabling zero-shot and few-shot adaptation to downstream tasks without explicit retraining. Such tuning typically optimizes for conditional likelihoods of task-specific outputs given multimodal inputs and natural language instructions, bridging the gap between foundational pretraining and practical application [33]. Importantly, the design of training curricula, sampling strategies, and data augmentation techniques also critically influence performance. Balancing modality-specific data scales, mitigating modality bias, and incorporating hard negative mining are active areas of research to improve robustness and fairness. Finally, considerations around computational cost and scalability shape training choices, as multimodal models often require orders of magnitude more parameters and data than unimodal counterparts. Understanding the landscape of training paradigms is thus essential for both advancing the theoretical foundations and enabling effective deployment of multimodal foundation models [34].

4 Cross-Modal Alignment and Fusion Techniques

A fundamental challenge in multimodal foundation models lies in effectively aligning and fusing information from disparate modalities such as vision, language, and audio, each characterized by distinct statistical properties, representational for-

mats, and semantic structures [35]. Cross-modal alignment refers to the process by which the model learns to map inputs from different modalities into a common representational space or to establish correspondences that enable coherent joint reasoning. Fusion, on the other hand, involves integrating these aligned representations to produce a unified understanding or generate multimodal outputs [36]. The design of alignment and fusion mechanisms profoundly influences the model's ability to generalize across tasks and domains, handle incomplete or noisy inputs, and exhibit compositional reasoning [37]. Formally, given modality-specific encoded representations $\{z^{(m)}\}_{m=1}^{M}$ with $z^{(m)} \in \mathbb{R}^{d_m \times T_m}$, where M is the number of modalities, d_m the feature dimension, and T_m the token or patch length, alignment typically aims to learn mappings or transformations $f_m : \mathbb{R}^{d_m \times T_m} \to \mathbb{R}^{d \times T}$ such that the resulting representations reside in a shared space $\mathbb{R}^{d \times T}$. This facilitates comparison, interaction, or fusion across modalities [38]. One canonical approach is to apply linear projection layers followed by normalization and contrastive objectives to align modality embeddings at the global or token level [39]. For example, in CLIP, global image and text embeddings $z^{(v)}$ and $z^{(t)}$ are projected via learned matrices W_v, W_t into a joint space, i.e.,

$$\tilde{z}^{(v)} = \frac{W_v z^{(v)}}{\|W_v z^{(v)}\|}, \quad \tilde{z}^{(t)} = \frac{W_t z^{(t)}}{\|W_t z^{(t)}\|}$$

enabling cosine similarity-based alignment [40]. While such global alignment suffices for retrieval tasks, finer-grained token-level or patch-level alignment is often necessary for generation and reasoning tasks, motivating cross-attention mechanisms. Cross-modal fusion frequently employs attention-based modules that enable dynamic information exchange between modalities [41]. Given query, key, and value matrices (Q, K, V) derived from different modality streams, cross-attention is computed as:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where Q may originate from one modality (e.g., text tokens) and (K, V) from another (e.g., image patches) [42]. This facilitates context-aware conditioning, allowing the model to selectively attend to relevant elements across modalities [43]. Architectures such as Flamingo and Uni-Perceiver extensively use interleaved or hierarchical cross-attention layers to integrate visual and textual features [44]. Beyond attention, fusion methods also include concatenation followed by feed-forward layers, bilinear pooling, tensor decompositions, and gating mechanisms that adaptively weigh modality contributions. Mixture-of-experts models further enhance fusion by routing inputs through modality-specialized subnetworks, improving scalability and specialization. Another important consideration is the

timing of fusion: early fusion integrates raw or low-level features before encoding, intermediate fusion merges latent representations within the network, and late fusion combines outputs or decisions from modality-specific heads. Each strategy presents trade-offs between expressiveness, efficiency, and robustness [45]. Early fusion can capture fine-grained interactions but often requires compatible input formats and high computational cost [46]. Late fusion is simpler and modular but may miss cross-modal contextual cues [47]. Intermediate fusion, particularly via cross-attention, has become a preferred approach in recent foundation models due to its flexibility and empirical effectiveness [48]. However, cross-modal alignment and fusion face multiple challenges including modality heterogeneity, varying token lengths, missing or noisy modalities, and semantic gaps [49]. Robustness to partial inputs and the ability to disentangle modality-specific from modality-agnostic factors remain active research frontiers [50]. Furthermore, interpretability of multimodal interactions and understanding how models fuse diverse information to produce coherent outputs is crucial for building trustworthy systems [51]. Future progress hinges on developing principled alignment objectives, scalable fusion architectures, and rigorous evaluation frameworks that capture multimodal reasoning capabilities beyond simple retrieval or generation [52].

5 Evaluation Benchmarks and Metrics for Multimodal Foundation Models

Evaluating multimodal foundation models presents unique challenges due to the diversity of modalities involved, the broad range of downstream tasks, and the complexity of measuring cross-modal understanding and reasoning [53]. Unlike unimodal models, where metrics such as perplexity for language models or accuracy for image classification provide relatively straightforward evaluation, multimodal models must be assessed on tasks that span retrieval, generation, classification, reasoning, and interaction across modalities [54]. Consequently, a rich ecosystem of benchmarks and metrics has emerged to comprehensively evaluate performance, generalization, robustness, and alignment with human expectations [55]. Formally, let $\mathcal{D} = \{(x_i^{(v)}, x_i^{(t)}, y_i)\}_{i=1}^{N}$ denote a multimodal test dataset, where $x_i^{(v)}$ and $x_i^{(t)}$ represent visual and textual inputs respectively, and y_i is the task-specific ground truth output [56]. Evaluation metrics $\mathcal{M}$ are designed to quantify the model's output $f_\theta(x_i^{(v)}, x_i^{(t)})$ against y_i [57]. For retrieval tasks, common metrics include Recall@K (R@K), which measures the proportion of queries for which the correct match is found within the top-K retrieved items, defined as

$$\mathrm{R@}K = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big(\mathrm{rank}(f_\theta(x_i^{(v)}), x_i^{(t)}) \leq K\Big)$$

where $\mathbf{1}$ is the indicator function and $\mathrm{rank}(\cdot)$ denotes the position of the true pair in the sorted retrieval list [58]. Metrics such as mean reciprocal rank (MRR) and normalized discounted cumulative gain (nDCG) further capture ranking quality. For generative tasks like image captioning, visual question answering (VQA), or multimodal machine translation, evaluation often relies on language generation metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE, which compare the generated text against reference captions or answers. However, these metrics have well-known limitations, especially regarding semantic adequacy and diversity, motivating research into embedding-based and learned evaluation metrics that measure semantic similarity in continuous spaces, such as BERTScore or CLIPScore [59]. Formally, embedding-based metrics compute similarity scores between generated outputs g_i and references r_i as

$$\mathrm{Score}(g_i, r_i) = \frac{\sum_{j=1}^{|g_i|} \max_k \cos\Big(\mathbf{e}(g_{i,j}), \mathbf{e}(r_{i,k})\Big)}{|g_i|}$$

where $\mathbf{e}(\cdot)$ denotes the embedding function, typically derived from pretrained language or multimodal models [60]. Robustness evaluation is also critical, given that multimodal models must handle noisy, incomplete, or adversarial inputs [61]. Metrics here include performance degradation under occlusions, modality dropout, or distribution shifts [62]. Evaluating fairness and bias, for instance measuring how model outputs vary with demographic attributes embedded in visual or textual content, is an emerging priority [63]. Metrics such as demographic parity difference or equalized odds adapted to multimodal settings are used to quantify such biases [64]. Multitask and zero-shot evaluation frameworks have become increasingly important, testing a model's ability to generalize without fine-tuning across tasks such as image classification, captioning, visual reasoning, speech recognition, and robotics control. Leaderboards like VQA, MSCOCO captioning, ImageNet zero-shot classification, and newer unified benchmarks such as HELM (Holistic Evaluation of Language Models) or MM-Vet provide standardized evaluation protocols. However, significant gaps remain in benchmarking compositionality, commonsense reasoning, and grounded interaction, which are crucial for real-world deployment. In summary, the evaluation landscape for multimodal foundation models is multifaceted and rapidly evolving. A combination of task-specific, embedding-based, robustness, and fairness metrics is necessary to holistically assess capabilities [65]. Designing benchmarks that reflect real-world complexity while remaining computationally tractable remains a vital research direction, as does the development of

interpretable and human-aligned evaluation methodologies that can guide future model improvements [66].

6 Challenges and Open Problems in Multimodal Foundation Models

Despite remarkable progress in building multimodal foundation models that integrate vision, language, audio, and other modalities, numerous fundamental challenges and open problems remain, impeding their broader adoption, robustness, and interpretability. These challenges arise from the inherent heterogeneity of modalities, the complexity of cross-modal interactions, and practical constraints such as data availability, computational cost, and ethical considerations. Addressing these issues is critical for advancing the theoretical foundations and real-world applicability of multimodal AI systems. One key challenge lies in the **modality gap**, which encompasses the semantic, statistical, and representational differences between modalities. Formally, given modality embeddings $z^{(m)} \in \mathbb{R}^{d_m}$ for each modality m, the distributional discrepancy $D(p_{z^{(v)}}, p_{z^{(t)}})$ between visual and textual latent spaces often leads to alignment difficulties, where D may be measured by metrics such as maximum mean discrepancy (MMD) or Wasserstein distance [67]. This gap complicates learning a unified latent space that faithfully captures joint semantics without losing modality-specific nuances [68]. Overcoming this requires innovative alignment methods that balance modality invariance and specificity while maintaining robustness to noisy or incomplete data. Another significant challenge is **scalability**. Multimodal models typically require massive datasets spanning multiple modalities, such as billions of image-text pairs or hours of aligned speech and video [69]. Training these models involves substantial computational resources, often beyond the reach of many research groups [70]. Additionally, the parameter counts of state-of-the-art models regularly exceed billions, exacerbating issues of energy consumption and environmental impact. Efficient training paradigms, parameter-efficient fine-tuning, and model compression techniques remain active areas of research aimed at democratizing multimodal AI[71]. **Generalization and robustness** are also pressing concerns [72]. Multimodal models frequently suffer from modality bias, where performance disproportionately depends on one dominant modality, potentially ignoring complementary signals [73]. This bias can be exacerbated by imbalanced training data distributions [74]. Robustness to domain shifts, adversarial perturbations, and missing modalities is still limited, constraining deployment in real-world noisy environments [75]. Formalizing robustness guarantees and developing principled approaches to modality dropout and uncertainty estimation are essential future directions [76]. In-

terpretability and explainability present additional open problems. Multimodal interactions involve complex nonlinear transformations and cross-attention mechanisms that are difficult to disentangle [77]. Understanding how specific modality inputs contribute to model decisions, especially in safety-critical applications such as medical diagnosis or autonomous systems, remains an open research frontier [78]. Techniques such as attention visualization, counterfactual explanations, and concept attribution adapted to multimodal settings require further refinement and standardization [71]. Ethical and societal challenges compound these technical issues [79]. Multimodal models can inadvertently learn and amplify biases present in large-scale training data, including stereotypes and discriminatory associations spanning text and imagery. Privacy concerns also arise when models are trained on sensitive or personal multimodal data [80]. Addressing these challenges demands rigorous auditing frameworks, bias mitigation algorithms, and transparent data governance policies. Mathematically, an overarching difficulty is the lack of unified theoretical frameworks to characterize multimodal representation learning. While unimodal learning benefits from well-established theories around generalization bounds, optimization landscapes, and information bottlenecks, extending these to the multimodal setting—with heterogeneous input spaces $\{\mathcal{X}^{(m)}\}_{m=1}^{M}$ and complex cross-modal losses $\mathcal{L} = \sum_m \alpha_m \mathcal{L}_m + \beta \mathcal{L}_{\text{cross}}$—remains largely open. Bridging this gap will illuminate principled design principles and training regimes. In conclusion, despite significant advances, multimodal foundation models face a constellation of interrelated challenges spanning modality alignment, scalability, robustness, interpretability, and ethics [81]. Addressing these open problems requires interdisciplinary efforts integrating theory, algorithm design, large-scale experimentation, and societal considerations to unlock the full potential of multimodal intelligence.

7 Future Directions and Emerging Opportunities

As multimodal foundation models continue to evolve and demonstrate transformative potential across a wide array of applications, the horizon reveals several promising future directions and emerging opportunities that can fundamentally advance the field. These directions not only address existing challenges but also open avenues for new capabilities, deeper understanding, and broader societal impact [82]. A key future direction is the development of **truly unified multimodal models** that seamlessly integrate an arbitrary number of modalities—including vision, language, audio, video, tactile signals, and even more abstract modalities such as knowledge graphs or sensor data—within a single coherent architecture. Moving beyond pairwise modality integration, such models aim to dynamically

accommodate and reason over heterogeneous input combinations. This necessitates flexible architectures capable of modality-agnostic processing and efficient fusion, potentially leveraging advances in modular networks, mixture-of-experts, and dynamic routing. Additionally, advances in self-supervised and few-shot learning will empower models to scale gracefully with limited labeled data for novel modalities or tasks. **Continual and lifelong multimodal learning** represents another vital opportunity. Real-world agents often encounter streams of multimodal data with evolving distributions and tasks. Designing foundation models that can incrementally incorporate new knowledge without catastrophic forgetting, adapt to shifting modalities, and autonomously discover novel concepts is crucial for building robust, adaptive AI systems [83]. Techniques such as meta-learning, replay buffers, and parameter-efficient fine-tuning tailored for multimodal contexts are likely to be pivotal in this endeavor [84]. Interpretable and explainable multimodal models will become increasingly important, especially in high-stakes domains such as healthcare, autonomous driving, and security [85]. Future research will focus on developing **transparent reasoning mechanisms** that reveal how information from each modality contributes to decisions, enabling users to trust and verify model outputs [86]. Integration of symbolic reasoning with neural architectures, causal inference methods, and interactive explanation interfaces are promising approaches that could enhance model accountability and user collaboration. Moreover, **multimodal foundation models have enormous potential to revolutionize human-computer interaction** by enabling more natural, immersive, and context-aware interfaces [87]. Advances in embodied AI, multimodal dialogue systems, and augmented reality can empower systems that perceive, understand, and respond to the full richness of human communication and environment. This requires models capable of grounding language in perception and action, as well as learning from multimodal interaction experiences. Ethical considerations and social impact will remain paramount in future work. Developing **responsible multimodal AI** involves not only mitigating bias and ensuring privacy but also designing inclusive datasets and fostering transparency in data collection and model deployment [88]. Collaborative frameworks involving stakeholders from diverse backgrounds will be essential to align multimodal AI development with societal values and norms [89]. From a theoretical standpoint, establishing **foundations for multimodal learning**—including rigorous generalization theories, robustness guarantees, and optimization analyses—will deepen understanding and guide principled model design [90]. Additionally, integrating multimodal foundation models with emerging paradigms such as causal learning, knowledge-enhanced reasoning, and neuro-symbolic AI presents fertile ground for research [91]. In summary, the future of multimodal foundation models is rich with challenges and unprecedented opportunities [92]. By pushing the bound-

aries of integration, adaptability, interpretability, and responsibility, the field is poised to unlock new frontiers in artificial intelligence that more fully capture the complexity and richness of the world's multimodal data [93].

8 Conclusion

Multimodal foundation models represent a paradigm shift in artificial intelligence, enabling the integration of diverse data modalities such as vision, language, audio, and beyond into unified, versatile architectures capable of solving complex, real-world problems. Throughout this survey, we have examined the rapid evolution of multimodal learning—from early fusion techniques and alignment strategies to state-of-the-art training paradigms, evaluation frameworks, and emergent challenges. The intricate interplay between modalities necessitates sophisticated architectural designs and objective formulations, which have driven significant advances in the fields of representation learning and cross-modal reasoning.

Despite these advances, the domain remains fraught with substantial open problems, including modality heterogeneity, scalability constraints, robustness under distributional shifts, and interpretability concerns. The ethical implications of large-scale multimodal models, particularly with respect to bias and privacy, underscore the critical need for responsible development practices and transparent evaluation methodologies. Addressing these multifaceted challenges demands interdisciplinary collaboration spanning machine learning, cognitive science, and human-computer interaction.

Looking forward, emerging opportunities in unified multimodal architectures, continual learning, explainability, and human-centric AI promise to propel the field toward increasingly powerful and trustworthy systems. Advances in theoretical understanding and novel application domains will further enrich the capabilities of multimodal foundation models, enabling more natural and effective communication between humans and machines.

In conclusion, while multimodal foundation models have achieved remarkable milestones, the path ahead is both challenging and exhilarating. Continued innovation and conscientious stewardship will be essential to harness their full potential and to create AI systems that are robust, interpretable, and beneficial across diverse societal contexts.

References

[1] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and

gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[2] Vandan Mujadia, Ashok Urlana, Yash Bhaskar, Penumalla Aditya Pavani, Kukkapalli Shravya, Parameswari Krishnamurthy, and Dipti Misra Sharma. Assessing translation capabilities of large language models involving english and indian languages. *arXiv preprint arXiv:2311.09216*, 2023.

[3] Anke Tang, Li Shen, Yong Luo, Yibing Zhan, Han Hu, Bo Du, Yixin Chen, and Dacheng Tao. Parameter efficient multi-task model fusion with partial linearization. *arXiv preprint arXiv.2310.04742*, 2023.

[4] Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[5] Yifan Chen, Devamanyu Hazarika, Mahdi Namazifar, Yang Liu, Di Jin, and Dilek Hakkani-Tur. Empowering parameter-efficient transfer learning by recognizing the kernel structure in self-attention. *arXiv preprint arXiv:2205.03720*, 2022.

[6] Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhensu Chen, Xiaopeng Zhang, and Qi Tian. Qa-lora: Quantization-aware low-rank adaptation of large language models. *arXiv preprint arXiv:2309.14717*, 2023.

[7] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Matthew Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[8] Yassine Zniyed, Thanh Phuong Nguyen, et al. Efficient tensor decomposition-based filter pruning. *Neural Networks*, 178:106393, 2024.

[9] Shuying Zhang, Jing Zhang, Hui Zhang, and Li Zhuo. Rastformer: region-aware spatiotemporal transformer for visual homogenization recognition in short videos. *Neural Computing and Applications*, 2024.

[10] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv.2402.09353*, 2024.

[11] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, 2022.

[12] Alan Ansell, E. Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavas, Ivan Vulic, and Anna Korhonen. Mad-g: Multilingual adapter generation for efficient cross-lingual transfer. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[13] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022.

[14] Rajas Chitale, Ankit Vaidya, Aditya Kane, and Archana Ghotkar. Task arithmetic with lora for continual learning. *arXiv preprint arXiv.2311.02428*, 2023.

[15] Dan Zhang, Sining Zhoubian, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*, 2024.

[16] Mushui Liu, Bozheng Li, and Yunlong Yu. Omniclip: Adapting clip for video recognition with spatial-temporal omni-scale feature learning. *arXiv preprint arXiv:2408.06158*, 2024.

[17] Sunghyeon Woo, Baeseong Park, Byeongwook Kim, Minjung Jo, Sejung Kwon, Dongsuk Jeon, and Dongsoo Lee. Dropbp: Accelerating fine-tuning of large language models by dropping backward propagation. *arXiv preprint arXiv:2402.17812*, 2024.

[18] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022.

[19] Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. Mv-adapter: Multimodal video transfer learning for video text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[20] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[21] Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*, 2023.

[22] Haixin Wang, Jianlong Chang, Yihang Zhai, Xiao Luo, Jinan Sun, Zhouchen Lin, and Qi Tian. Lion: Implicit vision prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[23] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. *arXiv preprint arXiv.2309.16039*, 2023.

[24] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[25] Weijieying Ren, Xinlong Li, Lei Wang, Tianxiang Zhao, and Wei Qin. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. *arXiv preprint arXiv:2402.18865*, 2024.

[26] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision*, pages 4894–4902, 2017.

[27] Yongchang Hao, Yanshuai Cao, and Lili Mou. Flora: Low-rank adapters are secretly gradient compressors. *arXiv preprint arXiv.2402.03293*, 2024.

[28] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[29] Shishuai Hu, Zehui Liao, and Yong Xia. Prosfda: prompt learning based source-free domain adaptation for medical image segmentation. *arXiv preprint arXiv:2211.11514*, 2022.

[30] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Actprompt: In-domain feature adaptation via action cues for video temporal grounding. *arXiv preprint arXiv:2408.06622*, 2024.

[31] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[32] Yi Zhang, Chun-Wun Cheng, Ke Yu, Zhihai He, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Node-adapter: Neural ordinary differential equations for better vision-language reasoning. *arXiv preprint arXiv:2407.08672*, 2024.

[33] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2020.

[34] Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. Vmt-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[35] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-lora: Serving thousands of concurrent lora adapters. *arXiv preprint arXiv:2311.03285*, 2023.

[36] Javier Gurrola-Ramos, Oscar Dalmau, and Teresa E Alarcón. A residual dense u-net neural network for image denoising. *IEEE Access*, 9:31742–31754, 2021.

[37] Yiwei Ma, Yijun Fan, Jiayi Ji, Haowei Wang, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. X-dreamer: Creating high-quality 3d content by bridging the domain gap between text-to-2d and text-to-3d generation. *arXiv preprint arXiv:2312.00085*, 2023.

[38] Zehao Xiao, Jiayi Shen, Mohammad Mahdi Derakhshani, Shengcai Liao, and Cees GM Snoek. Any-shift prompting for generalization over distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[39] Yaqing Wang, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models. *ArXiv*, abs/2205.12410, 2022.

[40] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[41] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 2022.

[42] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, abs/2101.00190, 2021.

[43] Haotong Qin, Xudong Ma, Xingyu Zheng, Xiaoyang Li, Yang Zhang, Shouda Liu, Jie Luo, Xianglong Liu, and Michele Magno. Accurate lora-finetuning quantization of llms via information retention. *arXiv preprint arXiv:2402.05445*, 2024.

[44] Bowen Zhao, Hannaneh Hajishirzi, and Qingqing Cao. Apt: Adaptive pruning and tuning pretrained language models for efficient training and inference. *arXiv preprint arXiv:2401.12200*, 2024.

[45] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Xingang Pan, and Bo Dai. Diffmorpher: Unleashing the capability of diffusion models for image morphing. *arXiv preprint arXiv:2312.07409*, 2023.

[46] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T. Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv.2312.12379*, 2023.

[47] Akshita Gupta, Gaurav Mittal, Ahmed Magooda, Ye Yu, Graham W Taylor, and Mei Chen. Losa: Long-short-range adapter for scaling end-to-end temporal action localization. *arXiv preprint arXiv:2404.01282*, 2024.

[48] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[50] Xi Wang, Laurence Aitchison, and Maja Rudolph. Lora ensembles for large language model fine-tuning. *arXiv preprint arXiv.2310.00035*, 2023.

[51] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, 2024.

[52] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 2020.

[53] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[54] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[55] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masqclip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[56] Pareesa Ameneh Golnari. Lora-enhanced distillation on guided diffusion models. *arXiv preprint arXiv:2312.06899*, 2023.

[57] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

[58] Yitao Liu, Chenxin An, and Xipeng Qiu. Y-tuning: An efficient tuning paradigm for large-scale pre-trained models via label representation learning. *Frontiers of Computer Science*, 18(4):184320, 2024.

[59] Shih-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard BW Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *The Twelfth International Conference on Learning Representations*, 2023.

[60] Yahao Hu, Yifei Xie, Tianfeng Wang, Man Chen, and Zhisong Pan. Structure-aware low-rank adaptation for parameter-efficient fine-tuning. *Mathematics*, 11(20):4317, 2023.

[61] Jiaxiang Cheng, Pan Xie, Xin Xia, Jiashi Li, Jie Wu, Yuxi Ren, Huixia Li, Xuefeng Xiao, Min Zheng, and Lean Fu. Resadapter: Domain consistent resolution adapter for diffusion models. *arXiv preprint arXiv:2403.02084*, 2024.

[62] Ning Ding, Yujia Qin, Guang Yang, Fu Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Haitao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juan Li, and Maosong Sun. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *ArXiv*, abs/2203.06904, 2022.

[63] Marinela Parovic, Goran Glavas, Ivan Vulic, and Anna Korhonen. Bad-x: Bilingual adapters improve zero-shot cross-lingual transfer. In *North American Chapter of the Association for Computational Linguistics*, 2022.

[64] Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu Lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*, 2023.

[65] Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[67] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[68] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*, pages 1022–1035, 2021.

[69] Duarte M. Alves, Nuno Miguel Guerreiro, João Alves, José Pombal, Ricardo Rei, José Guilherme Camargo de Souza, Pierre Colombo, and André F. T. Martins. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics*, pages 11127–11148, 2023.

[70] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[71] Yassine Zniyed, Thanh Phuong Nguyen, et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[72] Weisen Jiang, Baijiong Lin, Han Shi, Yu Zhang, Zhenguo Li, and James T. Kwok. Effective and parameter-efficient reusing fine-tuned models. *arXiv preprint arXiv.2310.01886*, 2023.

[73] Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. Differentiable prompt makes pre-trained language models better few-shot learners. *ArXiv*, abs/2108.13161, 2021.

[74] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[75] Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, and Tiejun Zhao. Loradrop: Efficient lora parameter pruning based on output evaluation. *arXiv preprint arXiv:2402.07721*, 2024.

[76] Yunqi Zhu, Xuebing Yang, Yuanyuan Wu, and Wensheng Zhang. Parameter-efficient fine-tuning with layer pruning on free-text sequence-to-sequence modeling. *arXiv preprint arXiv:2305.08285*, 2023.

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

[78] Dongshuo Yin, Yiran Yang, Zhechao Wang, Hongfeng Yu, Kaiwen Wei, and Xian Sun. 1% vs 100%: Parameter-efficient low rank adapter for dense predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[79] Jose Gallego-Posada, Juan Ramirez, Akram Erraqabi, Yoshua Bengio, and Simon Lacoste-Julien. Controlled sparsity via constrained optimization or: How I learned to stop tuning penalties and love constraints. In *Annual Conference on Neural Information Processing Systems*, 2022.

[80] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of

large language models. In *The Tenth International Conference on Learning Representations*, 2022.

[81] Aryo Pradipta Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. Parameter-efficient fine-tuning of llama for the clinical domain. *arXiv preprint arXiv:2307.03042*, 2023.

[82] Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[83] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[84] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on lora finetuning dynamics. *arXiv preprint arXiv:2406.08447*, 2024.

[85] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

[86] Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[87] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *arXiv preprint arXiv:2407.08674*, 2024.

[88] Seungwoo Yoo, Kunho Kim, Vladimir G Kim, and Minhyuk Sung. As-plausible-as-possible: Plausibility-aware mesh deformation using 2d diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4324, 2024.

[89] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016.

[90] Szymon Tworkowski, Konrad Staniszewski, Mikolaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Milos. Focused transformer: Contrastive training for context scaling. In *Advances in Neural Information Processing Systems*, 2023.

[91] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[92] Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, and Ying Shan. Low-rank approximation for sparse attention in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[93] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.