Tackling Multimodal Learning Challenges with Mixture-of-Expert: A Survey

Liangwei Zheng , **Wei Emma Zhang** , **Olaf Maennel** , **Lin Yue** and **Weitong Chen**
Adelaide University
{liangwei.zheng,wei.e.zhang,olaf.maennel,lin.yue,weitong.chen}@adelaide.edu.au

Abstract

Mixture-of-Experts (MoE) presents a naturally compatible and scalable framework for multimodal learning, demonstrating strong adaptability across diverse modalities and tasks. Despite its growing success, a comprehensive and systematic review on the MoE metho addressing multimodal challenges remains lacking. Existing surveys tend to evaluate either multimodal learning or MoE independently from method taxonomy, overlooking the unique interplay between them. This survey fills the gap by answering a central question: *How does MoE effectively resolve multimodal challenges?* We approach this from three key perspectives: (1) **MoE as an Efficient Multimodal Engine:** enabling scalable multimodal modeling by decoupling computational cost from parameter growth and mitigating modality redundancy through selective expert activation; (2) **MoE as a Multimodal Representation Learner:** integrating complementary multi-opinion expert knowledge to enrich alignment and interaction representations; and (3) **MoE as a Multimodal Adapter:** providing a modular and flexible mechanism to model imperfect data scenarios such as modality imbalance and missing modality. Through our extensive literature review, we identify critical research gaps, including interpretable routing, expert communication, modality integration, and lifelong multimodal learning. We position this survey as a foundation for future research toward interpretable and sustainable multimodal Mixture-of-Experts system.

1 Introduction

Real-world data modeling is a mixing of various heterogeneous data, also termed as multimodal data, which brings comprehensive and multiple perspectives understanding of target (e.g. a complete patient representation can be viewed as the composition of EHR, medical image, lab test and so on.) [Xu *et al.*, 2023]. In recent year, many works have made successful progress on multimodal learning in many applications such as medical [Zheng *et al.*, 2025; Jiang *et al.*, 2024] and generative AI [Feng *et al.*, 2023; Xue *et al.*, 2023]. We recognize four main objectives of multimodal learning from our literature reviews:

(1) Multimodal Model Scaling is an objective to efficiently scale the multimodal model and enable large learning capability without significant increase of computational cost. Multimodal dataset usually offers multiple modality perspective of one instance, which typically requires larger computational costs for one instance [Li *et al.*, 2024; Li *et al.*, 2025a].

(2) Fusion and Interaction is an objective to jointly learn the synthetic interaction when two or more modalities present in one instance. For example, combining visual and textual information for video sentiment analysis ensures that complementary signals are leveraged together. Such a fusion feature usually combines the unique information from each modality and generates an interacted feature from the combination [Lin *et al.*, 2024b; Yu *et al.*, 2024a].

(3) Alignment and Translation is an objective to map heterogeneous modalities into a shared representation space such that semantically attributes are closely matched. A well-aligned modality also contribute to translate the modality under heterogeneous structure. This enables cross-modal understanding, for example aligning spoken words with lip movements in video. The alignment objective usually emphasizes semantic consistency across modalities while preserve their structural discrepancy with shared information between heterogeneous modality [Mustafa *et al.*, 2022].

(4) Modality Robustness is an objective to build resilient modality representation under imperfect data stream, where the modality data can be missing, corrupted and incorrect. For example, severely missing chest ray modality for patient data while the EHR modality is often observed. It aims to build systems that exhibit dynamic resilience, ensuring that the model maintains a comprehensive understanding of the task despite fluctuations in data availability, quality, and the very composition of the modality stream [Zheng *et al.*, 2025; Huai *et al.*, 2025].

Many multimodal learning models recently are designed as dense and achieve acceptable performance. However, dense models face challenges on multimodal learning such as training inefficiency, poor generalization on learning multimodal heterogeneity structure and poor generalization on multi-task adaption [Baltrušaitis *et al.*, 2018]. To address achieve the four main challenges in multimodal learning, researchers

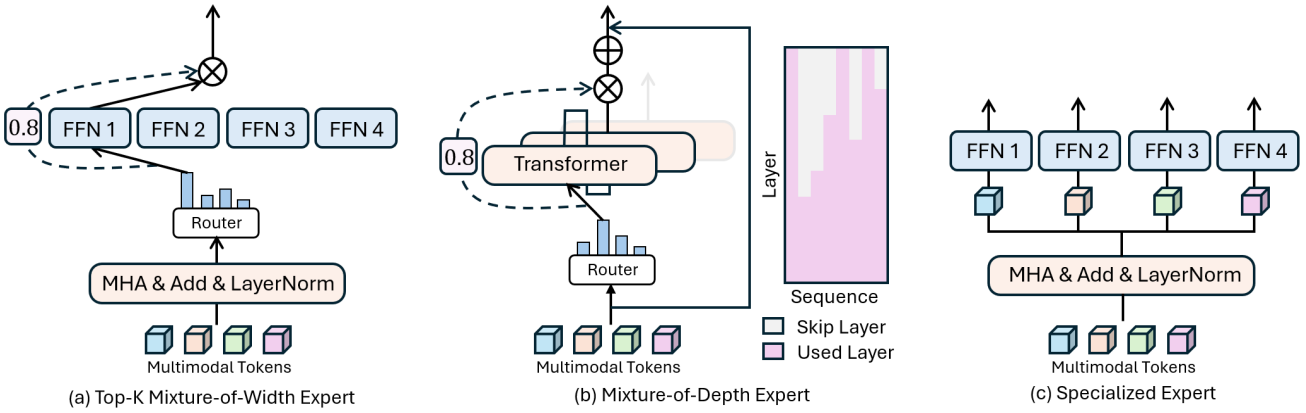

Figure 1: Types of Mixture-of-Experts for Multimodal Learning

found that Mixture-of-Experts (MoE) provides new learning paradigm that offers a natural solution to multimodal learning as shown in Figure 1: (1) MoE can be sparsely activated to scale up the multimodal model without significant increase of computational cost. (2) By routing modality-specific inputs to experts, localized representation learning that preserves modality-unique characteristics while facilitating cross-modal interaction. This specialization mechanism allows MoE to disentangle modality-dependent features and recompose them into semantically consistent representations. (3) MoE enables a multi-opinion mechanism that enriches the translation process. Instead of relying on a single shared representation, each expert captures unique perspectives of modality correspondence.

Despite the strong capability of MoE in multimodal learning, there still lack of a comprehensive literature reviews of relevant works to the best of our knowledge. There are some surveys discuss the multimodal learning [Xu *et al.*, 2023; Guo *et al.*, 2019] and MoE architecture [Masoudnia and Ebrahimpour, 2014; Cai *et al.*, 2025], but they somehow overlook that significance impacts and advantages of MoE for multimodal data and discuss only a few MoE methods in multimodal learning. In the light of this, we systematically study the most recent state-of-art (SOTA) methods as shown in Figure 2. Our taxonomy discuss the main challenges in multimodal learning and the specific MoE solution to address them. We primarily discuss (1) Multimodal scaling approach by the Mixture-of-Width and -Depth as in Figure 1. (2) Achieving multimodal alignment and interaction by expert specialized learning. (3) Addressing challenging multimodal problems by MoE such as missing modality and multimodal life-long learning. In addition, we further discuss current bottleneck and potential research directions for Multimodal MoE. We position this survey to bring complete views on process of using MoE for multimodal learning and to be specifically robust, so as to be an insightful guidebook for researchers who interest in multimodal MoE in this area.

2 Scope and Literature Collection Method

This survey collected papers from January 2020 to October 2025 from the following venues: ICLR, ICML, NeurIPS, ACL, EMNLP, CVPR, ECCV, ICCV, AAAI, IJCAI, KDD, WWW, MICCAI, and some are technical reports, and insightful preprint paper. Note that this survey will not discuss the papers where MoE is used simply as the feature extractor or simple replacement of FFN layer of Transformer without or with little technically and innovatively contribution to multimodal learning.

3 Related Survey

[Baltrušaitis *et al.*, 2018] starts a systematic review on multimodal learning, discussing applications and challenges on multimodal learning, providing representative multimodal works at the early stage of multimodal learning. [Guo *et al.*, 2019] further discussed on multimodal representation learning approach. [Xu *et al.*, 2023] revisited the history of Transformer for language, vision and multimodal learning from a topological perspectives and discuss the key components of Transformers in multimodal context in mathematical manner. More specifically, [Mai *et al.*, 2024] reviews multimodal learning for large language model, discussing key techniques such as multimodal chain of thought, multimodal instruction tuning, and multimodal in-context learning. In addition, [Mai *et al.*, 2024] highlighted fundamental and specific applications, modalities learning and design characteristics of multimodal LLM. [Jin *et al.*, 2024] further focus on efficient multimodal large model, considering MoE as one of the efficient training strategy, while focus more on other efficient techniques such as post-training, efficient attention mechanism, and compact multimodal architecture. [Han *et al.*, 2025] discuss generative model from test to any modality as in multimodal LLM, which categories six primary generative modalities and examines foundational techniques in multimodal generative model.

[Masoudnia and Ebrahimpour, 2014] systematically study the strength and challenges of MoE at the early development stage of MoE and [Cai *et al.*, 2025] bridge the essential gap of systematic review of literature on large language model with MoE structure, including the history, methods, applications and challenges in large language model.

Although there are many surveys conducting systematic reviews on either multimodal learning and MoE, most of

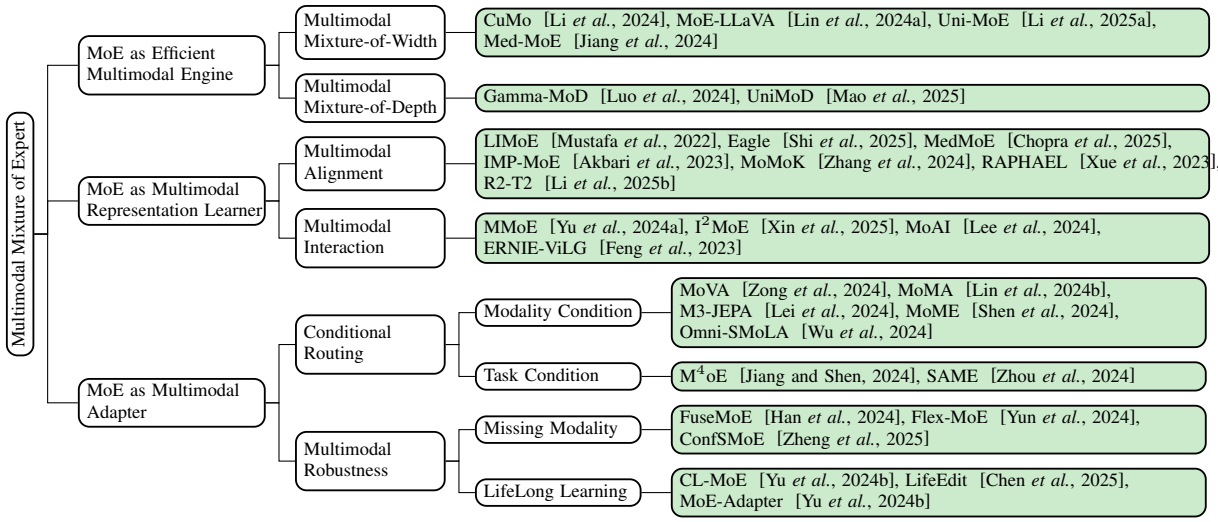

Figure 2: A taxonomy of approaches for multimodal learning with Mixture-of-Experts

surveys demonstrate limited discussion of MoE on multimodal learning, their discussion consider Mixture-of-Experts as only an efficient architecture and ignore the multi-views learning and the robust representation learning capability on multimodal learning. This overlook also underestimates the strength of MoE to robust modality adaption such as missing modality modeling and modality imbalance adaption.

4 Taxonomy of Multimodal Learning by Mixture of Expert

Our taxonomy systematically reviews Multimodal MoE through three branches: (1) MoE as Efficient Multimodal Engine: a category of scaling model by efficient MoE architecture including Mixture-of-Width Expert and Mixture-of-Depth Expert. (2) MoE as Multimodal Representation Learner: a category of learning strong interaction, aligned and fused multimodal representation by MoE characteristics. (3) MoE as Multimodal Adapter: a category of robust multimodal learning by MoE, including conditioning routing, missing modality modeling and lifelong learning. The main taxonomy is shown in Figure 2 and we provide details technical information of each approach with link to corresponding code repository if applicable in Table 1.

4.1 MoE as Multimodal Scaling Engine

The scalability of multimodal models is often constrained by the computational burden of processing heterogeneous, high-dimensional data [Li *et al.*, 2025a]. We categorize the scalable MoE into Mixture-of-Width Experts (MoWE) and Mixture-of-Depth Experts (MoDE). MoWE offers a solution by decoupling capacity from computation via sparse activation one or a few expert during forward, based on the selection criteria of router. MoDE, on the other hand, achieves scalability by skipping the forward layer of redundant token, which those tokens will be skip-connected to low-level layer or the output head.

Mixture-of-Width Expert. Upcycling [He *et al.*, 2024] is widely adopted to initialize experts from pre-trained dense models. In the multimodal context, this strategy is vital as it serves as a warm start that preserves established pre-trained capabilities while enabling specific experts to adapt to new modalities, effectively resolving the conflict between retaining general knowledge and learning cross-modal interactions. MoE-LLaVA [Lin *et al.*, 2024a] established a standardized pipeline for upcycling dense multimodal models. It adopts a three-stage training strategy: initially training the connector and fine-tuning the dense backbone, followed by a final stage where dense FFN layers are duplicated to initialize experts. This strategy allows the model to inherit pre-trained multimodal capabilities before specializing via sparse routing. While MoE-LLaVA focuses on upcycling the LLM backbone, CuMo [Li *et al.*, 2024] identifies redundancy in the vision encoder itself. It introduces Co-Upcycling, a strategy that simultaneously transforms both the CLIP vision encoder and the LLM's MLP layers into MoE architectures. This dual-sided sparsity enables finer-grained processing of visual patches alongside textual tokens. In addition, Uni-MoE [Li *et al.*, 2025a] generalizes upcycling techniques to "Any-to-Text" multimodal model by introducing more modality connectors, mapping diverse input modalities into the text space before integration with the LLM, showing the upcycling can effectively scale to manage heterogeneity of higher dimensional modalities. Med-MoE [Jiang *et al.*, 2024] further adapts this family of strategies to the medical domain. Building on the principles of CuMo and MoE-LLaVA, Med-MoE scales foundation multimodal medical models by treating experts within the MoE as domain specialists. Each expert is trained on a distinct medical dataset such as MRI and X-ray, reflecting the substantial variance across medical imaging modalities.

While standard approaches restrict upcycling to the pre-trained LLM backbone, recent works have expanded the technique's scope in two distinct dimensions. Uni-MoE applies upcycling to scale "Any-to-Text" modeling, validating MoE's

Method	Modality	Backbone Model	MoE Type	Task	Domain	Code
CuMo [Li *et al.*, 2024]	Image, Text	LLM & CLIP	Top-K MoWE	Multi-Task	General	
MoE-LLaVA [Lin *et al.*, 2024a]	Image, Text	LLM	Top-K MoWE	Multi-Task	General	
Uni-MoE [Li *et al.*, 2025a]	Image, Text, Audio, Speech, Video	LLM	Top-K MoWE	Multi-Task	General	
Med-MoE [Jiang *et al.*, 2024]	Image, Text	LLM	Top-K MoWE	Multi-Task	Medical	
Gamma-MoD [Luo *et al.*, 2024]	Image, Text	LLM	MoDE	Multi-Task	General	
UniMoD [Mao *et al.*, 2025]	Image, Text	LLM	MoDE	Multi-Task	General	
LIMoE [Mustafa *et al.*, 2022]	Image, Text	Transformer	Top-K MoWE	Representation Learning	General	
Eagle [Shi *et al.*, 2025]	Image, Text	LLM	Specialized Expert	VQA	General	
MedMoE [Chopra *et al.*, 2025]	Multimodal Brain Slice	Transformer	Specialized Expert	Representation Learning	Medical	N/A
IMP-MoE [Akbari *et al.*, 2023]	Image, Spectrogram, Text, Waveform	Transformer	Top-K MoWE	Representation Learning	General	N/A
MoMoK [Zhang *et al.*, 2024]	Image, Text, Knowledge Graph	Variational Network GNN	Specialized Expert	Representation Learning	Knowledge Graph	
RAPHAEL [Xue *et al.*, 2023]	Image, Text	Diffusion	Time Step Expert	Image Generation	General	N/A
R2-T2 [Li *et al.*, 2025b]	Image, Text	LLM	Top-K MoWE	Multi-Task	General	
MMoE [Yu *et al.*, 2024a]	Image, Text	LLM	Specialized Expert	Multi-Task	General	
I^2MoE [Xin *et al.*, 2025]	Image, Text, Time Series, Lab Test, Genetic, Bio-specimen	MoE	Specialized Expert	Multi-Task	Medical	
MoAI [Lee *et al.*, 2024]	Image, Text	LLM	Top-K MoWE	Multi-Task	General	
ERNIE-ViLG [Feng *et al.*, 2023]	Image, Text	Diffusion	Time Step Expert	Image Generation	General	
MoVA [Zong *et al.*, 2024]	Image, Text	LLM	Specialized Expert	Multi-Task	General	
Omni-SMoLA [Wu *et al.*, 2024]	Image, Text	VLM	Specialized Expert	VQA	General	N/A
MoMA [Lin *et al.*, 2024b]	Image, Text	Transformer	MoWE & MoDE	Multi-Task	General	N/A
M3-JEPA [Lei *et al.*, 2024]	Image, Text, Audio	MoE	Specialized Expert	Multi-Task	General	
MoME [Shen *et al.*, 2024]	Image, Text	LLM	Specialized Expert	Multi-Task	General	
M^4oE [Jiang and Shen, 2024]	Multitmodal Brain Slice	SwinUNet	Specialized Expert	Multi-Task	Medical	
SAME [Zhou *et al.*, 2024]	Image, Text	Transformer	Task-Aware Expert	Language Vision Navigation	General	
FuseMoE [Han *et al.*, 2024]	Chest Ray, ECG, Vital Signal, Text	Transformer	Top-K MoWE	Missing Modality Modeling	Medical	
Flex-MoE [Yun *et al.*, 2024]	Image, Text, Time Series, Lab Test, Genetic, Bio-specimen	Transformer	Top-K MoWE	Missing Modality Modeling	Medical	
ConfSMoE [Zheng *et al.*, 2025]	Time Series, Image, Text, ECG	Transformer	Top-K MoWE	Missing Modality Modeling	Medical	
CL-MoE [Huai *et al.*, 2025]	Image, Text	VLM	Top-K MoWE	Continual Learning	General	
LifeEdit [Chen *et al.*, 2025]	Image, Text	LLM	Specialized Expert	Model Editing	General	
MoE-Adapter [Yu *et al.*, 2024b]	Image, Text	VLM	Top-K MoWE	Continual Learning	General	

Table 1: A summary of Multimodal Mixture-of-Experts Approaches, organized according to the Taxonomy in Figure 2; Note that those paper with code not available up to January 2026

robustness in adapting to diverse multimodal data. In parallel, CuMo extends sparsity to the vision encoder itself, demonstrating that upcycling is broadly applicable to arbitrary pretrained modality encoders, not just language models.

Mixture-of-Depth Expert. Token redundancy describes the phenomenon where unnecessary tokens consume computational resources while contributing minimally to the model's final output [Mao *et al.*, 2025]. This inefficiency is particularly exacerbated in the multimodal setting, often referred to as the modality redundancy problem, due to the high density of visual tokens. To address this, Mixture-of-Depth-Experts (MoDE) frameworks introduce a dynamic mechanism to allocate computational budget by adjusting the network depth based on redundancy levels [Mao *et al.*, 2025]. In this paradigm, informative modalities (or tokens) are processed through the full-depth dense transformers, while redundant ones are routed to MoD pathways that strategically skip layers, thereby reducing computational overhead while preserving task performance. Gamma-MoD [Luo *et al.*, 2024] pioneers the application of Mixture-of-Depth to multimodal learning by identifying that modality redundancy is intrinsically linked to the low-rank structure of attention maps. By quantifying this via ARank (Attention Rank), it proposes a dual-path architecture: tokens exceeding a significance threshold are processed by a full depth Dense Transformer, while redundant tokens are offloaded to a Mixture-of-Depth (MoD) path for dynamic layer skipping. While Gamma-MoD treats redundancy as a static property of the input, UniMoD [Mao *et al.*, 2025] argues that token importance is highly context-dependent. Through empirical analysis of unified transformers, UniMoD reveals that redundancy is primarily dictated by the modeling objective—specifically,

generative tasks (e.g., diffusion) are far more sensitive to token loss than understanding tasks. It leverages distinct routers switched by task indices to dynamically adjust the pruning intensity based on ARank, ensuring that the model preserves necessary information density for generation while maximizing efficiency for understanding.

4.2 MoE as Representation Learner

MoE transcends its traditional role as a mere scaling tool, serving as a powerful engine for multimodal representation learning. Multimodal learning also benefits from the disentangled nature of expert to achieve modality specific and interactive learning. A single dense encoder often struggles to jointly capture semantic-level features while maintaining balanced representation quality across modalities [Yu *et al.*, 2024a]. Mixture-of-Experts (MoE) mitigates these issues by decomposing representation learning into specialized experts, each responsible for distinct attributes or modality-specific signals.

Multimodal Alignment. LIMoE [Mustafa *et al.*, 2022] pioneered the use of Sparse MoE as a unified representation encoder for multimodal data; by employing entropy-based regularization, it successfully balances expert usage across modalities, though it faces challenges in training stability due to the modality misbalance phenomenon. Addressing similar scalability issues, IMP-MoE [Akbari *et al.*, 2023] also utilizes a single sparse transformer but identifies that multi-loss learning (e.g., combining video, audio, text) generates conflicting gradients. IMP overcomes this by introducing Alternating Gradient Descent (AGD) to decouple updates, offering superior stability compared to LIMoE's standard optimization, albeit with a more complex training schedule. Shifting from internal sparse layers to external expert integration, Eagle [Shi *et al.*, 2025] explores a Mixture of Encoders design. Unlike LIMoE and IMP which train experts from scratch, Eagle leverages SOTA pre-trained vision encoders (e.g., for segmentation or detection) to achieve granular visual alignment with LLMs. While this yields high-resolution understanding, it creates a dependency on the quality of frozen external encoders. In the domain-specific arena, Med-MoE [Chopra *et al.*, 2025] and MoMoK [Zhang *et al.*, 2024] refine the expert concept for specialized reasoning. Med-MoE adopts a lightweight approach for medical VLMs, using a router to switch between domain experts (e.g., MRI, CT) and a meta-expert for shared knowledge, prioritizing resource efficiency and domain specific learning. MoMoK, focusing on Knowledge Graphs, innovates by explicitly minimizing the mutual information between experts. This disentanglement forces experts to learn distinct modality-specific knowledge, removing the blended expert often seen in standard MoEs, which an expert can be assigned with multimodal input. RAPHAEL [Xue *et al.*, 2023] employs a Space-MoE for spatial attribute binding and a Time-MoE for diffusion timesteps. This hierarchical structure ensures superior text-image alignment compared to standard diffusion models, though it incurs high computational complexity due to the billions of potential diffusion paths. R2-T2 [Li *et al.*, 2025b] defined the modality bias problem that routers in multimodal MoE models often assign suboptimal expert weights, limiting performance on unseen or complex tasks. R2-T2 proposed a test-time rerouting method that adjusts routing weights without retraining. The approach leverages a reference set of correctly predicted samples and modifies the routing weight of new test inputs by imitating their neighbors' routing patterns.

Multimodal Interaction. MMoE [Yu *et al.*, 2024a] fundamentally shifts multimodal fusion from a "one-size-fits-all" approach to an interaction-aware mechanism. Motivated by the observation that modalities interact differently, exhibiting redundancy (agreement), uniqueness (disagreement where one is correct), or synergy (combined efficacy), MMoE assigns input data to experts specialized in these specific interaction types. However, MMoE relies on fixed pseudo-labels derived from pre-trained model predictions to guide this routing, which limits scalability. Addressing this, I^2MoE [Xin *et al.*, 2025] extends the concept to arbitrary numbers of modalities using a loss-driven approach. Rather than relying on rigid pre-classification, I^2MoE learns to extract these interaction patterns dynamically through specialized fusion experts and a reweighting model, offering both scalability and interpretability in how modalities are prioritized. Taking a structural approach to interaction, MoAI [Lee *et al.*, 2024] explicitly engineers the mixing process. Instead of abstract interaction types, it introduces six dedicated attention experts handling specific pairwise flows (e.g., Visual-Language, Visual-Auxiliary). By leveraging outputs from external vision tools (segmentation, OCR) as auxiliary inputs, MoAI ensures fine-grained interaction between raw visual data, derived auxiliary knowledge, and language features, regulated by a lightweight gating network. In the generative domain, ERNIE-ViLG [Feng *et al.*, 2023] implements interaction through a Mixture-of-Denoising-Experts. Unlike the discriminative fusion in previous works, each expert here specializes in a specific diffusion time-step, fusing text and visual features to guide the generation process continuously.

4.3 MoE as Multimodal Adapter

Conditional Routing

Dense multimodal models often face limitations in specialized learning because a single encoder must simultaneously capture diverse modality signals and task-specific nuances, leading to diluted representations and suboptimal performance across tasks. Such models lack the flexibility to adapt contextually, since all inputs are processed through shared dense layers without distinction. MoE directly addresses this problem by assigning inputs to specialized experts that can focus on particular modalities. This specialization enables fine-grained feature extraction while maintaining computational efficiency through sparse routing.

Modality Condition. MoVA [Zong *et al.*, 2024] addresses the variability in visual tasks by treating Modality Condition as a routing problem among external encoders. Recognizing that no single vision encoder (e.g., CLIP, DINO) dominates all tasks, MoVA employs a coarse-to-fine strategy: it first uses an LLM to classify the input context (e.g., document vs. natural image) and then routes the image to the most suitable task-specific expert for fine-grained feature extraction. In contrast to MoME [Shen *et al.*, 2024], which also leverages multiple vision encoders but MoME argues that task in-

terference requires both visual and linguistic specialization. Therefore, it consider the expert as all vision encoders to aggregate multi-view features from all encoders simultaneously, constructing a comprehensive representation, then a Mixture of Language Experts (MoLE) adapts the LLM backbone using sparse, task-specific low-rank adapters. While MoVA and MoME focus on external encoder utilization, MoMA [Lin *et al.*, 2024b] and M3-JEPA [Lei *et al.*, 2024] bring modality conditioning into the internal model architecture. MoMA partitions the MoE layer into explicit modality groups (e.g., text experts vs. image experts). To ensure balanced training, it utilizes the Expert-Choice mechanism, where experts select tokens rather than tokens selecting experts, effectively handling the token-count disparity between modalities. It further leverages upcycling to initialize these sparse experts from dense checkpoints, accelerating convergence. M3-JEPA takes a distinct approach by operating in the latent prediction space rather than the generative space. It employs a Multi-Gate MoE as a shared predictor to map embeddings across modalities within a Joint-Embedding Predictive Architecture. Its gating mechanism explicitly disentangles modality-specific information from shared semantic content, optimizing contrastive alignment without the heavy reconstruction costs of generative models. Omni-SMoLA [Wu *et al.*, 2024] redefines the expert not as a feed-forward layer or an external encoder, but as a parameter-efficient fine-tuning module. It treats visual, text, and cross-modal interactions as separate domains, assigning each to a specialized LoRA expert. Unlike the hard routing in MoMA or MoVA, Omni-SMoLA employs a Soft-MoE mechanism, where tokens are softly mixed across these LoRA experts. This allows the model to blend visual and textual reasoning capabilities continuously, providing a flexible and efficient fine-tuning paradigm that avoids the catastrophic forgetting typical of generalist models.

Task Condition. Beyond modality-specific experts that focus on distinct input modalities, MoE can also be extended to task-specific experts, where each expert is dedicated to a particular multimodal task. This task condition expert can naturally address gradient conflict from multi-task objective as dense model [Liu *et al.*, 2021]. M^4oE [Jiang and Shen, 2024] leverage SoftMoE [Puigcerver *et al.*, 2023] technique to construct modality-specific MoE pool. One additional share modality MoE pool, but task specific, is adapted to construct task-guided feature, allowing the MoE to achieve dynamic modality fusion and modality-task dependence modeling for multi-task modeling. SAME [Zhou *et al.*, 2024] introduced task-embedding to apply over language-visual tokens for multimodal navigation task, allowing the experts in SAME to be aware task related feature from multimodal interaction.

Multimodal Robustness

Multimodal learning faces a dual challenge in real-world deployment: data is often structurally incomplete and temporally evolving [Zheng *et al.*, 2025; Wang *et al.*, 2020]. To achieve strong modality robustness, a model must not only handle the sudden absence of modality but also adapt its internal knowledge base to new tasks or privacy requirements over time. The MoE framework has emerged as a powerful paradigm for addressing these challenges. By decoupling a model's capacity from its computation through a sparse, modular architecture, MoE provides a natural mechanism for adaption of varying modality environment, where the experts can be designed to specific modality or combination. This modular architecture can easily adapt incomplete modality and manipulate the model knowledge.

Missing Modality. Fuse-MoE [Han *et al.*, 2024] construct all missingness pattern given a certain number of modality and each missingness pattern will be routed to one expert. FuseMoE also design Laplacian gate that routes missingness pattern in a non-extreme distribution and result in better load balance. Following a similar idea, Flex-MoE [Yun *et al.*, 2024] assigns each modality combination pattern to a designated expert. To ensure accurate routing, the router is optimized with a supervised loss that enforces correct expert selection. For incomplete samples, Flex-MoE introduces a learnable missing modality bank, which provides context-aware embeddings indexed by the observed modalities and the missing modality type. The completed representation is then routed to the corresponding expert, enabling specialized learning for that specific modality combination. ConfSMoE [Zheng *et al.*, 2025], on the other hand, adapts another imputation strategy: Two-stage imputation, incorporating the multi-perspective of expert as an important information for imputation. ConfSMoE first pre-imputed the missing modality by the mean of existing set from the missing modality, and then selectively post-imputed information from token-level cross-attention to existing modality after expert forward. In addition, ConfNet replaced the Softmax routing score with downstream task confidence, which address the gradient conflicts between router and load balance loss.

Life-Long Learning. LiveEdit [Chen *et al.*, 2025] implements a dynamic expert generator that produces instance-specific low-rank experts for VLLM editing, utilizing a dual-routing mechanism hard visual filtering and soft textual routing to integrate new data without disrupting foundational knowledge. MoE-Adapters [Yu *et al.*, 2024b] addresses the stability-plasticity trade-off in CLIP through its Distribution Discriminative Auto-Selector (DDAS), which preserves zero-shot capabilities by routing out-of-distribution inputs to the frozen backbone while directing task-specific data to specialized MoE adapters. Complementing these, CL-MOE [Huai *et al.*, 2025] manages non-stationary data in continual VQA via a hierarchical Dual-Router (RMoE) that optimizes expert selection at both task and instance levels, combined with a Momentum MoE (MMOE) that dynamically adjusts parameters to prevent catastrophic forgetting. Together, these methods demonstrate MoE's impact in decoupling model capacity from computation, allowing for the precise management of evolving multimodal information without the prohibitive costs of full-model retraining.

5 Findings and Future Directions

Throughout this comprehensive review, MoE models have demonstrated strong learning capability and scalability across diverse domains and tasks in multimodal learning. Nevertheless, several critical challenges remain insufficiently ad-

dressed, including interpretable routing mechanisms, effective expert communication, and life-long multimodal adaptation. Tackling these issues is crucial to enhance both the extendability and interpretability of multimodal MoE frameworks.

Interpretable routing mechanism for multimodal learning. Multimodal learning is well-known as heterogeneous structure [Luo *et al.*, 2024]. While many recent works such as CuMo [Li *et al.*, 2024], IMP-MoE [Akbari *et al.*, 2023], and Med-MoE [Jiang *et al.*, 2024], considers a hard Top-K routing mechanism without considering the heterogeneous structure of modality. For example, a modality can be too complicated to be learned by only one or a few expert, so as the Top-K is hard to identify the exact K for a modality, resulting hand-crafted finetuning and inefficient usage of resources. Although Gamma-MoD [Luo *et al.*, 2024] and Uni-MoD [Mao *et al.*, 2025] leveraged the *rank of attention map* (ARank) as a criteria determine the complexity of modality, ARank only focus on local metrics of one instance and overlook the global picture of modality. This can potential lead to instance bias. Addressing this problem can significantly increase the interpretability of multimodal MoE and further reduce computational effort.

Expert Communication and Modality Fusion. Most of works such as MMoE [Yu *et al.*, 2024a] and MoME [Shen *et al.*, 2024] consider MoE as specialized experts or mechanisms to scale up the model. These specialized experts hold their own "opinion" about one particular modality or one attribute of the data. However, the interaction between experts for multimodal learning remains limited exploration. One potential solution is to merge the knowledge of specialized experts into one global and comprehensive expert that integrates multi-view information and interactions from each expert. Such a unified expert could serve as a consensus learner, capable of capturing cross-expert dependencies, resolving inter-modality conflicts, and synthesizing higher-level semantic understanding across diverse modalities.

Limited and Simplified Multimodal Scope. Most existing works focus primarily on two modalities (e.g., image and text), such as MoVA [Zong *et al.*, 2024], RAPHAEL [Xue *et al.*, 2023], and MoE-LLaVA [Lin *et al.*, 2024a]. However, real-world data often comprise more than two modalities, and modeling only dual-modality interactions is insufficient to construct large-scale, generalizable foundation models capable of complex reasoning and multi-task adaptation. Moreover, real-world multimodal data frequently exhibit heterogeneous structures and modality missingness patterns, which challenge current dual-modality assumptions. Although models like I^2MoE [Xin *et al.*, 2025], Uni-MoE [Li *et al.*, 2025a], FuseMoE [Han *et al.*, 2024], and ConfS-MoE [Zheng *et al.*, 2025] extend to multiple modalities, they are either constrained by small-scale training or assume fully observed modalities, limiting their robustness and scalability in practical multimodal scenarios. The development of large-scale foundational models capable of handling both missing modalities and evolving data distributions remains an under-explored frontier in multimodal research.

Multimodal Efficient Expert Compression and Pruning Although MoE is designed to enhance computational efficiency by activating only a subset of experts, its effectiveness can deteriorate when the number of experts is not properly determined. In multimodal scenarios, modality heterogeneity introduces uneven feature complexity and representational demands, meaning that each modality may require a different hidden dimension and number of experts to achieve optimal learning. However, the estimation of suitable expert configurations within the feed-forward network remains insufficiently explored. Hence, establishing a principled criterion that reflects modality-specific complexity is necessary to guide the design of adaptive expert size and count, enabling more balanced computation, efficient parameter utilization, and stronger representational specialization across diverse modalities.

Modality Dependency and Evolving. A research [Wang *et al.*, 2020] finds that modality synergy, joint training often fails because different modalities generalize and overfit at inconsistent rates. This inconsistency is caused by modality imbalance, which the model becomes biased toward a dominant modality, leading to a performance drop where the joint network is outperformed by its best unimodal counterpart. In dynamic environments where modality dependencies shift over time, a MoE architecture can address this by evolving experts at the modality-dependent relationship. A properly designed, dependency-aware router can mitigate model bias by dynamically activating experts based on real-time modality relevance rather than architectural preference. This approach ensures that the model can adapt to temporal dependency shifts, maintaining the contributions of weak modalities even when strong ones dominate the initial learning phase.

6 Conclusion

In this survey, we present a systematic and comprehensive review of the literature on Multimodal MoE models, serving a valuable taxonomy for researcher exploring valuable contribution of MoE technologies to multimodal learning. We introduce new taxonomy specifically focusing on addressing multimodal challenges by MoE approach. Beyond the scaling property and strong representation learning capability of MoE, we highlight the practical challenges such as multimodal life-long learning, missing modality, modality imbalance problem in multimodal learning can also be addressed by MoE architecture. We hope this survey can contribute to an essential reference for researchers seeking to rapidly equip themselves with knowledge of multimodal MoE models and actively contribute to address critical challenges of multimodal learning.

Acknowledgments

This work was supported by Australian Research Council ARC Early Career Industry Fellowship (Grant No.IE240100275), Australian Research Council Discovery Projects (Grant No.DP240103070), Australian Research Council Linkage(Grant No.LP230200821), 2026 Global Partnership Joint Fund and Australia's Economic Accelerator Ignite (Grant No. IG250200014).

References

[Akbari *et al.*, 2023] Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam. Alternating gradient descent and mixture-of-experts for integrated multimodal perception. *Advances in Neural Information Processing Systems*, 36:79142–79154, 2023.

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[Cai *et al.*, 2025] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Chen *et al.*, 2025] Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, and Xiaofeng He. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9455–9466, 2025.

[Chopra *et al.*, 2025] Shivang Chopra, Gabriela Sanchez-Rodriguez, Lingchao Mao, Andrew J Feola, Jing Li, and Zsolt Kira. Medmoe: Modality-specialized mixture of experts for medical vision-language understanding. *arXiv preprint arXiv:2506.08356*, 2025.

[Feng *et al.*, 2023] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023.

[Guo *et al.*, 2019] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019.

[Han *et al.*, 2024] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37:67850–67900, 2024.

[Han *et al.*, 2025] Longzhen Han, Awes Mubarak, Almas Baimagambetov, Nikolaos Polatidis, and Thar Baker. Multimodal large language models: A survey. *arXiv preprint arXiv:2506.10016*, 2025.

[He *et al.*, 2024] Ethan He, Abhinav Khattar, Ryan Prenger, Vijay Korthikanti, Zijie Yan, Tong Liu, Shiqing Fan, Ashwath Aithal, Mohammad Shoeybi, and Bryan Catanzaro. Upcycling large language models into mixture of experts. *arXiv preprint arXiv:2410.07524*, 2024.

[Huai *et al.*, 2025] Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19608–19617, 2025.

[Jiang and Shen, 2024] Yufeng Jiang and Yiqing Shen. M4oe: A foundation model for medical multimodal image segmentation with mixture of experts. In *international conference on medical image computing and computer-assisted intervention*, pages 621–631. Springer, 2024.

[Jiang *et al.*, 2024] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *arXiv preprint arXiv:2404.10237*, 2024.

[Jin *et al.*, 2024] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024.

[Lee *et al.*, 2024] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models. In *European Conference on Computer Vision*, pages 273–302. Springer, 2024.

[Lei *et al.*, 2024] Hongyang Lei, Xiaolong Cheng, Dan Wang, Kun Fan, Qi Qin, Huazhen Huang, Yetao Wu, Qingqing Gu, Zhonglin Jiang, Yong Chen, et al. M3-jepa: Multimodal alignment via multi-directional moe based on the jepa framework. *arXiv preprint arXiv:2409.05929*, 2024.

[Li *et al.*, 2024] Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:131224–131246, 2024.

[Li *et al.*, 2025a] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Li *et al.*, 2025b] Zhongyang Li, Ziyue Li, and Tianyi Zhou. R2-t2: Re-routing in test-time for multimodal mixture-of-experts. *arXiv preprint arXiv:2502.20395*, 2025.

[Lin *et al.*, 2024a] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

[Lin *et al.*, 2024b] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024.

[Liu *et al.*, 2021] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.

[Luo *et al.*, 2024] Yaxin Luo, Gen Luo, Jiayi Ji, Yiyi Zhou, Xiaoshuai Sun, Zhiqiang Shen, and Rongrong Ji. $\gamma-$ mod: Exploring mixture-of-depth adaptation for multimodal large language models. *arXiv preprint arXiv:2410.13859*, 2024.

[Mai *et al.*, 2024] Xinji Mai, Zeng Tao, Junxiong Lin, Haoran Wang, Yang Chang, Yanlan Kang, Yan Wang, and Wenqiang Zhang. From efficient multimodal models to world models: A survey. *arXiv preprint arXiv:2407.00118*, 2024.

[Mao *et al.*, 2025] Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unimod: Efficient unified multimodal transformers with mixture-of-depths. *arXiv preprint arXiv:2502.06474*, 2025.

[Masoudnia and Ebrahimpour, 2014] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

[Mustafa *et al.*, 2022] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[Puigcerver *et al.*, 2023] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

[Shen *et al.*, 2024] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070, 2024.

[Shi *et al.*, 2025] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *The Thirteenth International Conference on Learning Representations, 2025, Singapore*, 2025.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[Wu *et al.*, 2024] Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14205–14215, 2024.

[Xin *et al.*, 2025] Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *arXiv preprint arXiv:2505.19190*, 2025.

[Xu *et al.*, 2023] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[Xue *et al.*, 2023] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36:41693–41706, 2023.

[Yu *et al.*, 2024a] Haofei Yu, Zhengyang Qi, Lawrence Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 2024*. Association for Computational Linguistics, 2024.

[Yu *et al.*, 2024b] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024.

[Yun *et al.*, 2024] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805, 2024.

[Zhang *et al.*, 2024] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. *arXiv preprint arXiv:2405.16869*, 2024.

[Zheng *et al.*, 2025] Liangwei Nathan Zheng, Wei Emma Zhang, Mingyu Guo, Miao Xu, Olaf Maennel, and Weitong Chen. Rethinking gating mechanism in sparse moe: Handling arbitrary modality inputs with confidence-guided gate. *arXiv preprint arXiv:2505.19525*, 2025.

[Zhou *et al.*, 2024] Gengze Zhou, Yicong Hong, Zun Wang, Chongyang Zhao, Mohit Bansal, and Qi Wu. Same: Learning generic language-guided visual navigation with state-adaptive mixture of experts. *arXiv preprint arXiv:2412.05552*, 2024.

[Zong *et al.*, 2024] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 37:103305–103333, 2024.