MoE-ViE: Mixture of Experts Vision Encoder for Efficient Image and Video Understanding

Bonan Zhang[1], Shiyu Dong[1], Quan Hung Tran[1], Katharina Gschwind[⋆], Shuqi Yang[⋆], Sijia Chen[⋆], Adel Ahmadyan[1], Seungwhan Moon[1], Lu Zhang[1], Ahmed Kirmani[1], Babak Damavandi[1], and Anuj Kumar[1]

[1]Meta
[⋆]Work done while at Meta
{bonanz,shiyudong,quanhungtran,kgschwind,shuqiyang
,sijiac,ahmadyan,luzhang20,kirmani,babakd,anujk}@meta.com

Abstract. Vision encoders are a critical component of vision-language models, and scaling their capacity effectively improves performance. However, dense scaling increases compute cost and inference latency. Mixture-of-Experts (MoE) architectures offer a compelling alternative, having enabled efficient scaling in LLMs, yet the MoE design space for CLIP-style vision encoders remains underexplored at State-of-the-Art (SOTA) levels. In this work, we systematically study MoE designs for vision encoder scaling and find that fine-grained MoE topologies yield substantial gains over both dense and standard MoE counterparts. We further propose an auxiliary-loss-free balancing variant for better expert utilization, and design a specialized MoE kernel to mitigate inference latency overhead. To enhance video capabilities while preserving image knowledge, we introduce frame-level distillation paired with a novel freezing mechanism. We pretrain a series of Mixture-of-Experts Vision Encoders (MoE-ViE) across a range of sizes, all consistently outperforming their dense counterparts. Our largest model matches the zero-shot performance of a SOTA encoder 1.7× its size at 76% of its latency. When aligned with an LLM, MoE-ViE surpasses all compared encoders on image and video benchmarks, including those with up to 5× more activated parameters.

Keywords: Vision encoder · Mixture-of-Experts · CLIP

1 Introduction

Vision encoders have become a cornerstone in today's vision-language models (VLMs). The pursuit of higher accuracy and broader applicability necessitates the continuous scaling up of vision models, which in turn has been a reliable route to stronger VLM capabilities [5, 11, 25, 64]. However, scaling dense vision encoders comes with system-level trade-offs [10, 18]. Increasing their capacity typically induces growth in compute cost and inference-time latency. This issue becomes even more acute for high-resolution images or long videos, where moderate increases in encoder cost compound into substantial end-to-end overhead.

Mixture-of-Experts (MoE) offers an alternative scaling mechanism by increasing the total model capacity while only activating a small subset of experts for any given input, thus decoupling model capacity from compute cost [58]. MoE has been widely used in large language models (LLMs) [1,33,44,66,79] and has carried over to VLMs [3,15,28,42,65]. In contrast, applying MoE to vision encoders, particularly to the prevailing contrastive language-image pretraining (CLIP) [55], remains less settled. While prior research provides several attempts on developing MoE vision encoders, none has closed the gap to State-of-the-Art (SOTA) performance on core vision benchmarks [40,51,57,73,85].

In this work, we present a systematic study of MoE architecture design in vision encoders across multiple scales. We observe that conventional MoE [58], *i.e.*, treating MoE as a drop-in replacement for dense MLP blocks, provides only limited gains. In comparison, MoE vision encoders with fine-grained expert designs achieve larger, more consistent improvements. This aligns with insights from recent work on MoE-based LLMs [44]. MoE with fine-grained expert design is also well matched to the nature of encoding vision information, where a vision encoder is responsible for supporting a wide variety of visual concepts, styles, and domains. We further propose a variant of loss-free balancing [71] to balance expert utilization during training. One practical challenge is the system overhead introduced by memory-bound operations in MoE computations, which can hurt the latency benefits implied by sparsely activated parameters. We therefore develop an MoE kernel in Triton [68] to overcome this.

Beyond these architectural choices, we focus on refining our training strategy to emphasize building a unified image and video encoder. We first pretrain on a vast set of image-text pairs via contrastive training and then, following [5], we finetune with video data to enhance the model's capability in video understanding. However, we observe that vanilla video finetuning can shift representations in ways that degrade our pretrained image capabilities. We mitigate this by introducing a frame-level distillation during video finetuning and by freezing MoE experts. Together, we are able to preserve the image representation learned during pretraining while allowing video supervision to strengthen video representation. Fig. 1 summarizes our MoE architecture and training pipeline.

Our investigation yields several key findings:

- We show that with appropriate architectural design, vision encoders with MoE consistently outperform their dense counterparts across multiple scales. This confirms the efficacy of sparse scaling for CLIP-style vision encoders.
- We demonstrate that the MoE design proposed in this work substantially improves over prior works on MoE vision encoders across model sizes.
- Our trained MoE-ViE achieves SOTA results at all scales on zero-shot image and video benchmarks. Our largest encoder achieves results comparable to the SOTA dense encoder PE_{core}G, which is 1.7× larger.
- Our MoE kernel provides $> 2.5\times$ speedup in inference latency. It can be used as a drop-in implementation for different MoE topologies.
- Aligned with an LLM, MoE-ViE yields strong downstream results, and matches or exceeds other SOTA encoders in comparable settings.

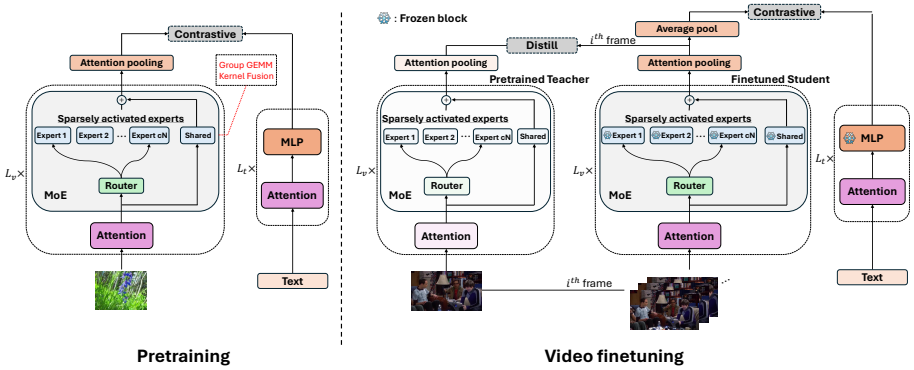


Fig. 1: Illustration of MoE-ViE architecture and training pipeline. MoE-ViE adopts a fine-grained MoE design with optimized MoE kernels for latency reduction. We first perform contrastive pretraining on large-scale image-text pairs. During video finetuning, we introduce frame-level distillation and expert freezing to preserve the pretrained image understanding capabilities.

2 Vision Encoder with MoE

In this section, we first review the MoE formulation, then present our studies on introducing MoE into CLIP vision encoders, including the architectural design and the control of expert utilization.

2.1 MoE Preliminaries

A typical MoE layer replaces the dense MLP block with N experts $\{E_i\}_{i=1}^{N}$. Given an input token $x \in \mathbb{R}^d$, a router produces logits

$$s = W_r x \in \mathbb{R}^N \tag{1}$$

followed by selecting a subset of experts via top-k routing. The standard Softmax-based gating computes

$$g(x) = \text{Softmax}(\text{top-}k(s)) \tag{2}$$

Denote $\mathcal{T}(x) \subset \{1, \ldots, N\}$ to be indices of selected experts, the MoE output is

$$y = \sum_{e \in \mathcal{T}(x)} g_e(x) E_e(x) \tag{3}$$

MoE increases the parameter count with N while keeping the per-token compute proportional to k experts, providing a favorable capacity vs. compute trade-off when $k \ll N$.

2.2 MoE for Vision Encoder

Adapting MoE to CLIP vision encoders raises three key design questions: (i) what expert topology is most appropriate for vision representations; (ii) how to ensure balanced expert utilization; (iii) how to realize the theoretical efficiency gains promised by MoE. We address these questions in the following sections.

Architecture Design Prior CLIP MoE works replace each dense MLP with N experts whose capacity matches the original MLP [40, 51, 73, 85]. We instead adopt a fine-grained expert design, *i.e.*, each expert has reduced hidden width relative to the original dense MLP, enabling more experts under a comparable compute budget. This choice is motivated by the heterogeneous nature of visual features (*e.g.*, color, shape, spatial layout) and by findings from LLMs that finer expert granularity yields more efficient scaling [17]. We apply MoE only to the vision tower and keep the text tower unchanged, since only the vision tower is used in any subsequent VLM alignment.

In addition to specialized experts, we include m shared experts that are always active, inspired by the idea of retaining a global thumbnail alongside high-resolution tiles in [45] as well as the success in LLMs [17]. The shared experts provide a persistent pathway for global context, while the routed experts capture conditional, token-specific transformations. We demonstrate expert specialization in more detail in Section 4.5.

We further replace the standard MoE Softmax gating with a Sigmoid function to avoid potential expert competition [16, 52]. Given logits s, we compute[1]

$$g(x) = \text{top-}k(\text{Sigmoid}(s)) \tag{4}$$

$$g'_e(x) = \frac{g_e(x)}{\sum_{i \in \mathcal{T}(x)} g_i(x)} \tag{5}$$

where $g'(x)$ renormalizes the selected scores to keep the mixture scale stable.

Let there be cN experts in total, where each expert has hidden width reduced by c. We reserve the first m experts as shared and route over the remaining specialized experts, *i.e.*, $\mathcal{T}(x) \subset \{m{+}1, \ldots, cN\}$. The layer output is

$$y = \sum_{e \in \mathcal{T}(x)} g'_e(x) E_e(x) + \lambda \sum_{e=1}^{m} E_e(x) \tag{6}$$

where hyperparameter λ controls the weight of shared experts. We set $\lambda = 1$ in all experiments for simplicity, as our model is not sensitive to this choice. Additional details are provided in Appendix A.2.

Table 1 compares MoE designs trained on 12.8B samples from MetaCLIP [77]. We observe consistent improvements from the introduction of MoE, and the fine-grained design leads to stronger performance.

[1] Some works, *e.g.*, [57], switch the order of Softmax and top-k selection to enable gradient flow in top-1 scenario, which is not needed for Sigmoid. For consistency with recent MoE models [39, 44], we also adopt this switch.

Table 1: Comparison of different MoE designs in vision encoders.

			General Classification							Image Retrieval				
Model	Total params	Activated params	Avg. Class.	ImageNet Val [19]	ImageNet v2 [56]	ObjectNet [4]	ImageNet Adversarial [30]	ImageNet Renditions [29]	Avg. Retrieval	COCO T→I [43]	COCO I→T [43]	Flickr30k T→I [81]	Flickr30k I→T [81]
ViT-B/32 dense	0.1B	0.1B	55.8	67.5	59.3	50.6	27.8	73.7	58.4	36.3	54.4	63.3	79.6
ViT-B/32 conventional MoE	0.5B	0.1B	59.9	70.6	62.3	54.3	34.7	77.7	59.7	37.8	56.3	64.4	80.2
MoE-ViE-B/32	0.5B	0.1B	**63.2**	**72.9**	**64.9**	**57.7**	**40.4**	**80.0**	**61.4**	**38.9**	**56.6**	**66.5**	**83.5**
ViT-L/16	0.3B	0.3B	78.0	79.7	72.6	74.9	72.4	90.5	68.9	45.8	63.8	75.9	90.3
ViT-L/16 conventional MoE	1.7B	0.3B	78.5	80.2	72.9	73.5	74.9	90.8	69.0	45.5	63.4	76.1	90.8
MoE-ViE-L/16	1.7B	0.3B	**79.6**	**80.7**	**74.1**	**76.5**	**75.4**	**91.5**	**69.3**	**46.0**	**63.8**	**76.3**	**91.0**

Load Balancing Compared to prior works designing various auxiliary losses [51, 57], optimized jointly with the main training objective, we propose a variant of the loss-free approach [71], which decouples the control of expert utilization from the training objective. In loss-free balancing, an extra bias term b is introduced to the router. Given routing logits s, the router now selects experts by

$$g(x) = \text{top-}k(\text{Sigmoid}(s) + b) \tag{7}$$

During training, the bias for expert e is updated using the observed token load. Let t_e be the number of tokens routed to expert e in the current iteration, and let μ_t denote the mean token count across experts. The update rule is

$$b_e = b_e - \alpha \ \text{sign}(t_e - \mu_t) \tag{8}$$

with step size α. We notice, however, that the sign$(\cdot)$ update applies a constant magnitude correction (*e.g.*, 1 or -1) regardless of deviation, which can cause oscillations when routing is close to balanced. Therefore, we propose to replace sign$(\cdot)$ with z-score, which scales the correction by the degree of imbalance and keeps it normalized for stability. The update of the bias term then becomes

$$b_e = b_e - \alpha \frac{t_e - \mu_t}{\sigma_t} \tag{9}$$

where σ_t is the standard deviation of $\{t_e\}$. We present ablations on different load balancing strategies in Section 4.4.

Scalability We then explore the scalability of MoE-ViE under contrastive pre-training. We train our models at multiple scales. At each scale, we train both the baseline dense model and MoE-ViE. All models are trained for 12.8B samples on MetaCLIP [77] at 224×224 resolution, and evaluated using ImageNet-1K classification accuracy [19]. Fig. 2 summarizes the results. Across all scales, MoE-ViE consistently outperforms its dense counterpart under the same compute budget, indicating a desired model capacity vs. compute trade-off. This motivates us to further scale up MoE-ViE as detailed in Section 4.1.

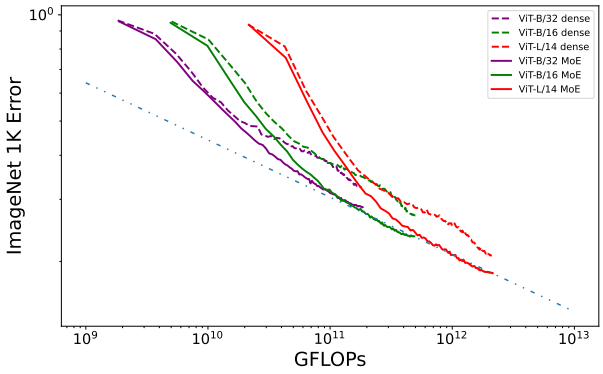


Fig. 2: MoE-ViE vs. dense models at multiple scales. MoE-ViE consistently outperforms their dense counterparts under the same compute budget.

2.3 Latency Optimization

Inherent MoE Efficiency Unlike dense networks where FLOPs scale linearly with total parameter count, MoE maintains constant active FLOPs per token by routing inputs to a small subset of experts. Thus, MoE theoretically offers superior representation (via more parameters) without an inherent latency penalty, as its compute cost is governed only by the number of active parameters.

Optimized Kernel While MoE is theoretically compute-efficient, naive implementations (*e.g.*, sequentially looping over experts in PyTorch) often underutilize GPUs, mainly due to three factors: (1) fragmenting compute into many small general matrix multiply (GEMM) operations reduces arithmetic intensity; (2) dynamic routing introduces CPU–GPU synchronization to determine per-expert workloads, creating idle hardware "bubbles"; (3) larger intermediate tensors amplify HBM traffic, making memory bandwidth the latency bottleneck.

To realize MoE's theoretical efficiency in practice, we implement a specialized Triton [68] kernel with two key optimizations: First, *Grouped GEMM* aggregates all expert matrix multiplications (MatMuls) into a single, unified GPU task, enabling parallel processing across experts and eliminating repeated CPU–GPU handshakes. Second, *Kernel Fusion* combines MatMuls and non-linear activations (*e.g.*, GeLU) into one pass, minimizing round-trips to HBM and reducing data movement. More details are included in Appendix B.

These optimizations are not auxiliary architectural tweaks. Instead, they are the necessary implementation mechanisms to faithfully translate MoE's algorithmic efficiency into practical, hardware-efficient execution.

3 Robust Video Finetuning

Following [5], we first pretrain the model on image-text pairs and then finetune on video-text pairs to build a unified vision encoder for both image and video

understanding. However, in our experiments, naively finetuning on video data substantially degrades image performance, indicating severe forgetting. Mixing image data into the video stage partially recovers image accuracy, but we find it also limits gains on video understanding (Section 4.4).

Recognizing that most VLMs use frame-based vision encoders to encode videos frame by frame [13,80], we regularize the video finetuning with frame-level distillation. We keep a copy of the image-pretrained model as a teacher. For a given video, we randomly sample a frame and feed it through the student and teacher vision towers to obtain logits S and T, and minimize the cosine distance

$$\mathcal{L}_d = 1 - \cos(S, T) \tag{10}$$

The final loss becomes

$$\mathcal{L} = \mathcal{L}_c + \beta \mathcal{L}_d \tag{11}$$

where $\mathcal{L}_c$ denotes contrastive training loss, and β is the weight of distillation.

We further reduce forgetting by freezing the MoE experts during video finetuning. This aligns with findings that MLP layers store models' acquired knowledge, while attention layers are primarily responsible for feature interaction [21]. Since our vision encoder operates on individual frames, we do not expect video finetuning to require large changes to the per-frame visual vocabulary. Instead, the adaptation should mainly update how frame features are aggregated for video-specific alignment.

One issue is that $\mathcal{L}_c$ updates both towers. So when finetuning on videos, the text tower may drift away from the alignment established during pretraining, while $\mathcal{L}_d$ encourages the vision tower to stay close to its image-pretrained teacher, creating an optimization mismatch. We find that an efficient way to eliminate this is to freeze the MLP layers in the text tower as well. Our intuition is that the text tower has already trained on substantially larger and more diverse data, and our primary goal in this stage is to adapt the vision encoder to videos.

4 Experiments

We develop a series of MoE-ViE at various scales by building on top of the standard CLIP vision encoder architecture, where we replace the MLP modules in the vision tower with our proposed MoE blocks. We exclude the first transformer block from this replacement, as this first layer captures general, low-level, visual information, and we did not observe any gains by specializing it with MoE in our experiments.

4.1 Zero-Shot Benchmarks

Setup We conduct contrastive pretraining for MoE-ViE on 3.5B image-text pairs, comprised of 2B data from MetaCLIP [77], and 1.5B proprietary data. Unless otherwise specified, all of our MoE variants use 32 experts for each MoE layer to maintain a practical compute budget. Each expert is designed to be 1/4

of the corresponding dense MLP. We include more studies on the impact of the number of experts in Appendix C.

We first train MoE-ViE-B and MoE-ViE-L models with 1/8 of experts activated, and then train an MoE-ViE-H with 1/4 experts activated to explore further scaling. Following [5], we perform progressive resolution training, *i.e.*, we increase image resolution as training progresses, for better training efficiency. After pretraining, we perform a short warmup using the pretraining data augmented with CC12M [7] and TreeOfLife [27]. For video finetuning, we uniformly sample 8 frames per video and use the curated video-text data from [5, 83]. Full training and evaluation details are provided in Appendix D.1 and Appendix D.2, respectively. Additional LLM alignment results are reported in Appendix E.

Video Benchmarks We report the zero-shot video understanding capability of MoE-ViE in Table 2 on both classification and retrieval tasks. MoE-ViE achieves SOTA results at all evaluated scales, and even outperforms the larger PE_{core}G model, which indicates that MoE scaling transfers effectively from image pretraining to our proposed video finetuning.

Table 2: Comparison of zero-shot video benchmarks between MoE-ViE and other previous SOTA vision encoders in the literature.

				Video Classification					Video Retrieval		
Model	Total params	Activated params	Resolution	Avg. Class.	K400 [35]	K600 [35]	UCF101 [63]	HMDB [38]	Avg. Retrieval	MSR-VTT T→V [78]	MSR-VTT V→T [78]
SigLIP2-B/16 [69]	0.1B	0.1B	224	59.5	58.7	55.0	82.0	42.3	34.3	38.5	30.1
PE_{core}B/16 [5]	0.1B	0.1B	224	65.9	65.6	65.1	**84.6**	48.2	47.5	47.6	**47.3**
MoE-ViE-B/16	0.5B	0.1B	224	**67.5**	**68.3**	**65.6**	84.5	**51.5**	**47.6**	**47.9**	47.2
SigLIP2-L/16 [69]	0.3B	0.3B	384	66.0	65.3	62.5	86.7	49.3	36.5	41.5	31.4
PE_{core}L/14 [5]	0.3B	0.3B	336	72.9	73.4	**72.7**	87.1	**58.5**	**50.2**	50.3	**50.1**
MoE-ViE-L/16	1.7B	0.3B	384	**73.4**	**74.5**	71.8	**89.2**	57.9	**50.2**	**50.5**	49.8
InternVL-C/16 [12]	5.5B	5.5B	224	-	69.1	68.9	-	-	42.5	44.7	40.2
SigLIP2-g-opt/16 [69]	1.1B	1.1B	384	69.8	69.8	67.0	90.7	51.8	38.7	43.1	34.2
PE_{core}G/14 [5]	1.9B	1.9B	448	76.2	**76.9**	**75.1**	90.7	61.1	**50.6**	51.2	**49.9**
MoE-ViE-H/14	3.5B	1.1B	448	**76.5**	**76.9**	75.1	**92.8**	**61.3**	**50.6**	**51.6**	49.5

General Image Benchmarks Table 3 summarizes zero-shot performance on image classification and retrieval benchmarks. We compare MoE-ViE to both SOTA dense vision encoders and to prior CLIP MoE vision encoders. MoE-ViE achieves the best results at each scale among models with comparable sizes. In particular, MoE-ViE demonstrates consistently high accuracy on the challenging ImageNet-A [30] benchmark, with +5.6% over PE_{core}B, +0.2% over PE_{core}L,

and even +0.6% over $PE_{core}G$ which has 1.7× as many activated parameters as MoE-ViE-H.

Table 3: Comparison of zero-shot image classification and retrieval between MoE-ViE and other SOTA vision encoders in the literature.

				General Image Classification							Image Retrieval				
Model	Total params	Activated params	Resolution	Avg. Class.	ImageNet Val [19]	ImageNet v2 [56]	ObjectNet [4]	ImageNet Adversarial [30]	ImageNet Renditions [29]	Avg. Retrieval	COCO T→I [43]	COCO I→T [43]	Flickr30k T→I [81]	Flickr30k I→T [81]	
LIMOE-B/16 [51]	0.5B	0.1B	224	-	73.7	-	-	-	-	-	36.2	51.3	-	-	
CLIP-UP-B/16 [73]	0.5B	0.2B	224	-	76.9	-	-	-	-	74.2	**52.1**	**71.5**	**80.9**	92.3	
SigLIP2-B/16 [69]	0.1B	0.1B	224	74.0	78.2	71.4	73.6	55.0	**91.7**	73.7	52.1	68.9	80.7	93.0	
$PE_{core}B/16$ [5]	0.1B	0.1B	224	74.6	78.4	71.7	71.9	62.4	88.7	74.3	50.9	71.0	80.8	**94.4**	
MoE-ViE-B/16	0.5B	0.1B	224	**76.8**	**79.3**	**72.8**	**74.0**	**68.0**	89.9	**74.4**	52.0	70.7	**80.9**	93.9	
CLIP-MoE-L/14 [85]	0.9B	0.5B	336	-	74.6	68.5	33.5	-	-	53.6	46.8	65.0	42.1	60.5	
LIMOE-L/16 [51]	1.7B	0.3B	224	-	78.6	-	-	-	-	-	39.6	55.7	-	-	
CLIP-UP-L/14 [73]	1.7B	0.5B	224	-	81.2	-	-	-	-	75.5	53.9	73.8	82.0	92.4	
SigLIP2-L/16 [69]	0.3B	0.3B	384	85.0	83.1	77.4	84.4	84.3	**95.7**	76.7	55.3	71.4	85.0	95.2	
$PE_{core}L/14$ [5]	0.3B	0.3B	336	86.1	83.5	**77.9**	84.7	89.0	95.2	**78.8**	57.1	**75.9**	85.5	**96.6**	
MoE-ViE-L/16	1.7B	0.3B	384	**86.2**	**83.6**	**77.9**	**84.8**	**89.2**	95.4	**78.8**	**57.2**	**75.9**	**85.7**	96.5	
EVA 18B/14 [64]	17.5B	17.5B	224	83.6	83.8	77.9	82.2	87.3	95.7	77.5	56.2	73.6	83.3	96.7	
InternVL-C/16 [12]	5.5B	5.5B	224	82.5	83.2	77.3	80.6	83.8	95.7	78.6	58.6	74.9	85.0	95.7	
SigLIP2-g-opt/16 [69]	1.1B	1.1B	384	88.0	85.0	79.8	88.0	90.5	96.6	77.6	56.1	72.8	**86.0**	95.4	
$PE_{core}G/14$ [5]	1.9B	1.9B	448	**88.6**	**85.4**	**80.2**	**88.2**	92.6	**96.5**	**78.9**	**58.1**	**75.4**	85.7	**96.2**	
MoE-ViE-H/14	3.5B	1.1B	448	88.3	85.1	80.0	87.0	**93.2**	96.2	78.2	56.8	74.6	85.4	96.0	

Fine-grained Image Benchmarks Table 4 presents the results for image classification on fine-grained categories. We see MoE-ViE excels in these tasks. Not only do our MoE-ViE models achieve SOTA results compared with similarly sized models, here, again, our largest model MoE-ViE-H outperforms $PE_{core}G$, whose activated size is 1.7× larger. Additionally, our model demonstrates high accuracy on OCR benchmarks, as evaluated on TextCaps [60]. We attribute these achievements to the nature of MoE, applied with the correct topology, enabling experts to specialize in different tasks.

4.2 VLM Alignment

Since VLMs represent the primary downstream application of vision encoders, evaluating alignment quality is essential to verify that improvements in visual representations translate to gains on end-user tasks such as visual question answering, video understanding, and captioning. Our alignment training follows a standard three-stage pipeline. In Stage 1 (warmup, 8k steps), we train only the two-layer MLP projectors while keeping all other parameters frozen. In Stage 2 (pre-alignment, 15k steps), we explore various training strategies, and find that freezing the encoder parameters for the first 1k iterations to stabilize the loss

Table 4: Comparison of zero-shot fine-grained image classification and OCR benchmarks between MoE-ViE and SOTA dense vision encoders. We re-evaluated part of the OCR results for SigLIP2 and PE if not reported in [69] and [5].

				Fine-grained Img Classification							**OCR**		
Model	Total params	Activated params	Resolution	Avg. Class.	Food101 [6]	Flowers [53]	Country211 [67]	Aircrafts [46]	Cars [37]	Average OCR	Text Cap T→I [60]	Text Cap I→T [60]
SigLIP2-B/16 [69]	0.1B	0.1B	224	69.2	92.8	85.7	19.2	54.8	**93.4**	71.5	72.3	70.7
PE_{core}B/16 [5]	0.1B	0.1B	224	71.7	92.5	86.5	30.5	57.0	92.1	71.6	72.3	70.9
MoE-ViE-B/16	0.5B	0.1B	224	**73.5**	**94.2**	**87.3**	**34.9**	**58.2**	93.1	**72.5**	**72.7**	**72.3**
SigLIP2-L/16 [69]	0.3B	0.3B	384	76.1	96.1	**90.0**	31.6	67.0	**95.8**	79.2	**80.2**	78.2
PE_{core}L/14 [5]	0.3B	0.3B	336	78.1	96.2	87.2	45.6	**67.8**	93.7	79.2	79.8	78.5
MoE-ViE-L/16	1.7B	0.3B	384	**78.9**	**96.7**	89.2	**47.8**	65.7	94.9	**79.7**	80.0	**79.3**
EVA 18B/14 [64]	17.5B	17.5B	224	75.9	95.8	86.0	43.1	59.7	94.9	-	-	-
InternVL-C/16 [12]	5.5B	5.5B	224	72.8	95.3	85.8	35.1	53.3	94.4	-	-	72.3
SigLIP2-g-opt/16 [69]	1.1B	1.1B	384	79.6	97.0	91.5	40.1	73.6	**95.9**	79.8	80.3	79.2
PE_{core}G/14 [5]	1.9B	1.9B	448	83.8	96.9	91.4	**57.6**	78.2	94.7	79.1	79.3	78.8
MoE-ViE-H/14	3.5B	1.1B	448	**83.9**	**97.1**	**92.4**	54.3	**80.1**	95.5	**80.4**	**80.6**	**80.2**

before transitioning to end-to-end training of all parameters yields the best downstream performance. In Stage 3 (supervised finetuning, 8k steps), we train all parameters of the model. Results are reported in Table 5. We use the tile-based approach [14, 72] to support dynamic resolutions.

Table 5: Alignment comparison across models. Columns are grouped by task type. Llama 3.1 Instruct 8B [26] is used as the base LLM, and we use a maximum of 4 tiles.

		Image						**Video**				**Captioning**		
Model	Activated params	Avg. Images	AI2D [36]	TextVQA [62]	ChartQA [48]	DocVQA [50]	Info. QA [49]	Avg. Video	VideoMME [24]	MVBench [41]	EgoSchema [47]	Avg. Cap.	COCO [43]	NoCaps [2]
MetaCLIP-G/14 [77]	1.8B	61.3	72.8	65.4	68.1	61.3	39.1	45.4	46.5	44.7	45.0	125.5	134.4	116.5
SigLIP2-g-opt/16 [69]	1.1B	59.0	72.4	70.3	63.1	55.3	34.0	49.5	46.2	48.5	53.8	129.1	137.8	120.3
PE_{core}G/14 [5]	1.9B	69.2	69.7	74.3	73.4	81.2	47.6	48.6	46.0	48.7	51.2	123.7	134.5	112.9
InternViT2.5/14 [11]	5.5B	68.1	72.9	71.3	**74.6**	74.3	47.6	48.7	46.0	49.6	50.6	123.0	132.5	113.5
AIMv2 3B/14 [23]	2.7B	69.8	72.2	**79.2**	73.0	78.2	46.5	49.7	49.6	49.9	49.6	**130.6**	**139.7**	**121.5**
MoE-ViE-H/14	1.1B	**75.2**	**84.9**	75.7	74.2	**86.1**	**55.0**	**58.9**	**52.0**	**63.6**	**61.2**	127.5	135.4	119.6

As shown in Table 5, MoE-ViE-H achieves strong performance across all evaluation axes. On both image and video benchmarks, our model outperforms all baselines within the same parameter class and even surpasses SOTA models with several times more activated parameters. In captioning, although we do not achieve the highest scores, our results remain highly competitive: among

models with fewer than 2B activated parameters, MoE-ViE-H is second only to SigLIP2-g-opt. We provide more analysis on alignment in Appendix E.2.

4.3 Latency Comparison

With our kernel optimization presented in Section 2.3, we achieve inference latency comparable to a dense model (SigLIP2-g-opt) with a similar number of active parameters, despite possessing a significantly larger pool of trainable parameters. As shown in Table 6, the MoE kernel provides $> 2.5\times$ speedup across different batch sizes compared to the vanilla implementation. Our MoE-ViE runs at roughly 76% of latency compared to $\text{PE}_{\text{core}}\text{G}/14$, the $1.7\times$ larger SOTA dense vision encoder, while demonstrating comparable performance on accuracy. This confirms that the representational gains of our proposed MoE-ViE are achieved without any unexpected latency regressions. Additionally, we further demonstrate that such efficiency gain is naturally translated to downstream VLMs, detailed in Appendix E.3.

Table 6: Performance comparison between MoE-ViE and dense vision encoders, across different inference batch sizes (*bsz*). Due to difference in patch size, we report latency based on the same number of input tokens (# tokens = 576) for a fair comparison.

Model	Activated Parameters	Total Parameters	Latency (ms)			
			***bsz*=16**	**32**	**64**	**128**
SigLIP2-g-opt/16 [69]	1.1B	1.1B	76.39	141.78	275.94	549.62
$\text{PE}_{\text{core}}\text{G}/14$ [5]	1.9B	1.9B	101.21	188.76	366.23	708.62
MoE-ViE-H/14 (Vanilla Implementation)	1.1B	3.5B	318.76	448.93	740.84	1306.81
MoE-ViE-H/14 (Optimized Kernel)	1.1B	3.5B	82.59	145.82	276.87	544.96

4.4 Ablation Study

Auxiliary Loss Table 7 compares different expert balancing algorithms. We conduct these experiments with MoE-ViE-B with patch size 32, which is faster to train, but exhibits the same routing dynamics as other MoE-ViE variants. As seen, the loss-free approaches (*i.e.*, [71] and ours) outperform other forms on all benchmarks. This aligns with our expectation, since expert utilization is achieved without directly perturbing optimization under the contrastive training objective. Our proposed magnitude-aware method further improves over [71], indicating a more stable and effective balancing control.

Video Finetuning We ablate the video finetuning components introduced in Section 3 in Fig. 3, using MoE-ViE-H. For clarity, we report ImageNet-1K as a representative image benchmark and K400 as a representative video benchmark.

Table 7: Comparison of different auxiliary losses for training CLIP vision encoders.

	General Classification							Image Retrieval				
Auxiliary loss	Avg. Class.	ImageNet Val [19]	ImageNet v2 [56]	ObjectNet [4]	ImageNet Adversarial [30]	ImageNet Renditions [29]	Avg. Retrieval	COCO T→I [43]	COCO I→T [43]	Flickr30k T→I [81]	Flickr30k I→T [81]
Dense model	55.8	67.5	59.3	50.6	27.8	73.7	58.4	36.3	54.4	63.3	79.6
Importance and load loss [57]	59.9	70.6	62.3	54.3	34.7	77.7	59.7	37.8	56.3	64.4	80.2
Entropy loss [51]	59.9	70.7	62.2	53.9	35.5	77.1	60.0	38.5	56.4	65.0	80.1
Loss-free [71]	62.6	72.5	64.4	56.4	40.0	79.9	60.9	**39.2**	**56.8**	65.3	82.5
Ours	**63.2**	**72.9**	**64.9**	**57.7**	**40.4**	**80.0**	**61.4**	38.9	56.6	**66.5**	**83.5**

As shown in Fig. 3a, while naive video finetuning on video data improves video performance, it causes a significant degradation on image benchmarks over time. Mixing image data into the video finetuning partially mitigates this issue, but fails to prevent degradation, and limits our gains on video understanding. We find that applying frame-level distillation without freezing the MLP layers in the text tower sharply hurts image performance. Finally, we see that, individually, both distillation and expert freezing help preserve image performance, and that their combination achieves the best results on both image and video benchmarks.

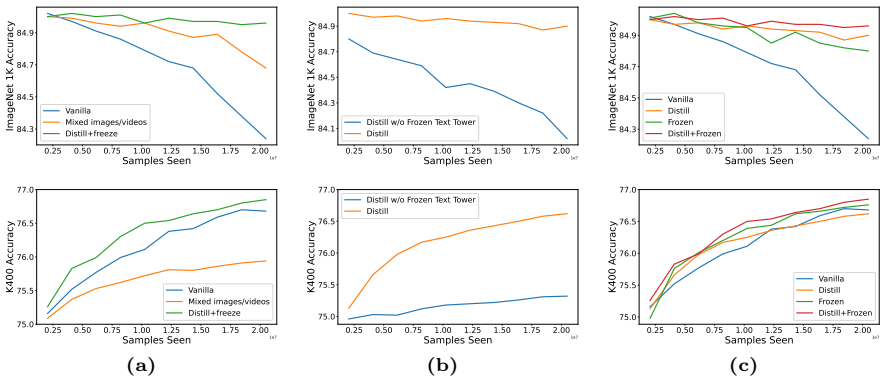
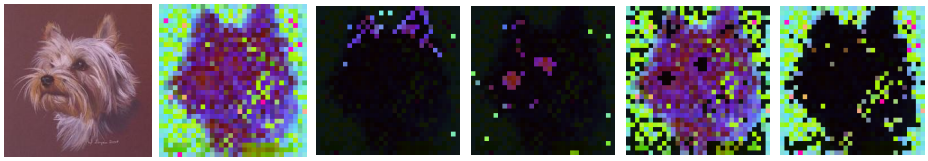
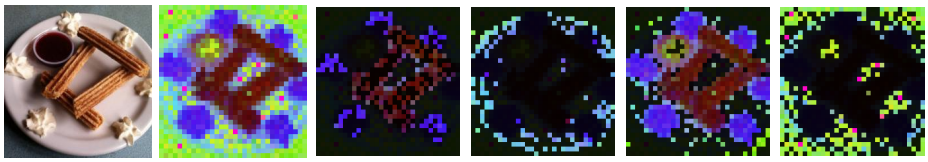


Fig. 3: Ablations of different video finetuning methods. (a) Our proposed distillation and freezing techniques outperform finetuning with vanilla video data, or mixed image and video data; (b) We observe significant degradation if we apply frame-level distillation without freezing the text tower; (c) The combination of the proposed distillation and freezing methods yields the best results.



(a) From left to right: raw image, shared expert, experts 11, 18, 22, 29



(b) From left to right: raw image, shared expert, experts 11, 17, 19, 29

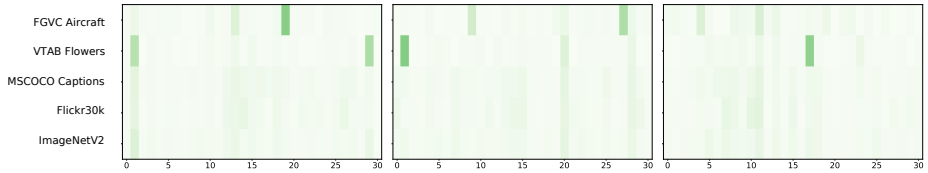
Fig. 4: MoE Expert activations in layer 20 of MoE-ViE-H/14. Experts specialize for distinct, semantically meaningful, aspects of the image.

4.5 Expert Specialization

We conduct both qualitative and quantitative analysis of MoE-ViE's expert activation patterns to better understand how fine-grained MoE captures complex, multi-faceted visual information. Across both tokens and tasks, we observe clear and meaningful expert specialization.

We first project each layer's features to three principal components using PCA, and map the resulting 3D embeddings to RGB, as in [5, 61]. For each expert, we visualize the tokens routed to it, as shown in Fig. 4. Since we use top-k routing, these subplots are not mutually exclusive, *i.e.*, a token may appear in multiple expert-specific views. We also illustrate the subplot for the shared expert, which receives all tokens, preserving global context.

Such visualization reveals several compelling insights: (1) each expert extracts distinct, semantically meaningful aspects of the image; (2) these specializations are complementary, such that their combined outputs yield richer visual understanding. For example, in Fig. 4a, expert 11 predominantly activates on tokens corresponding to the dog's ears, while expert 18 emphasizes the eyes and nose. In Fig. 4b, expert 17 highlights the plate, whereas expert 19 concentrates on the food. In both examples, expert 29 consistently activates on background regions.



(a) MoE layer 29 **(b)** MoE layer 30 **(c)** Last MoE layer, layer 31

Fig. 5: Expert activation across all data for a given task, final layers of MoE-ViE-H/14.

We further quantify expert specialization through per-task expert utilization heatmaps (Fig. 5) and routing entropy (Table 8). As shown in Fig. 5, experts exhibit distinct activation profiles across tasks, consistent with the emergence of task-relevant specialization in deeper MoE layers. Interestingly, on domain-specific benchmarks (*e.g.*, FGVC Aircraft, VTAB Flowers), activations are more imbalanced across experts than on more general benchmarks such as ImageNet or COCO. This concentration may help explain MoE-ViE's stronger performance on fine-grained image benchmarks. To complement this analysis, we measure routing entropy across tasks and layers, summarized in Table 8. Entropy consistently decreases with depth, confirming higher specialization at later layers. Moreover, benchmarks spanning narrower visual domains (VTAB Flowers, FGVC Aircraft) exhibit lower entropy overall, indicating stronger semantic specialization.

Table 8: Routing entropy averaged over layer ranges.

Benchmark	L0–5	L6–15	L16–25	L26–30	Uniform
VTAB Flowers [53]	3.20	3.17	2.83	2.30	3.47
FGVC Aircraft [46]	3.19	3.26	2.95	2.56	3.47
ImageNetV2 [56]	3.34	3.31	3.30	3.21	3.47
Flickr30k [81]	3.28	3.30	3.23	3.07	3.47
COCO [43]	3.30	3.32	3.30	3.18	3.5

5 Related Works

5.1 Contrastive Language-Image Pretraining

Contrastive vision-language pretraining has emerged as a foundational approach for building strong and transferable vision encoders [32, 55]. A number of extensions refine either its training objective or supervision signals. For instance, SigLIP [84] replaces the Softmax used to compute InfoNCE loss [54] with a Sigmoid function, enabling better scaling and efficiency. MaskCLIP [20] augments masked image modeling to train the vision model as a self-distillation. CoCa [82] incorporates caption loss in contrastive training, and LocCa [70] further introduces grounding supervision. SigLIP 2 [69] and PE [5] integrate several of these ingredients, achieving high accuracy on zero-shot image and video benchmarks.

5.2 Mixture-of-Experts (MoE)

MoE scales model capacity by replacing dense layers with an expert router and a set of expert layers. The learned expert router activates a small subset of experts per input, yielding sparse input-dependent computation. The idea traces back to conditional computation with expert gating formulations [8, 31, 34], and has since become a practical mechanism for scaling transformers via sparse top-k

routing [58]. MoE has been extensively studied in natural language processing (NLP) [22,39], and has shown great success in many recent large language models [1,33,44,66,79]. Early MoE for vision efforts largely follow supervised training strategies [9,57,75] or introduce MoE during the post-training stages [59,74,76]. Later works have introduced MoE into CLIP pretraining [55]. LIMoE [51] pioneered contrastive learning with sparsely activated experts and demonstrated improved zero-shot image benchmarks. CLIP-MoE [85] and CLIP-UP [73] introduce upcycling techniques to convert pretrained dense CLIP models to MoE variants. Despite these advances, prior MoE vision encoders have not matched performance of SOTA dense models.

6 Conclusion

In this work, we present MoE-ViE, an efficient and performant vision encoder with fine-grained MoE, for image and video understanding. By combining our proposed architectural design, kernel optimization, magnitude-aware loss-free load balancing control, and refined video finetuning, MoE-ViE outperforms both prior dense and MoE vision encoders across image and video benchmarks at all scales. Our largest model achieves comparable results to a SOTA dense encoder that is 1.7× larger while running at 76% of its latency, demonstrating the significant potential for further effectively scaling up vision encoders.

Acknowledgements

We would like to thank Giang Nguyen, Jiankun Liu, Zhuangqun Huang, Tuyen Tran, Xiao Zhang, Parisa Hassanzadeh, Lambert Mathias, Po-Yao Huang, Daniel Bolya for their contributions, support and discussions for this project.

References

1. Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R.K., Bai, Y., Baker, B., Bao, H., et al.: gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925 (2025)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)

5. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H.A., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li, S.W., Dollar, P., Feichtenhofer, C.: Perception encoder: The best visual embeddings are not at the output of the network. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=INqBOmwIpG
6. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
7. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
8. Chen, K., Xu, L., Chi, H.: Improved learning algorithms for mixture of experts in multiclass classification. Neural networks **12**(9), 1229–1252 (1999)
9. Chen, T., Chen, X., Du, X., Rashwan, A., Yang, F., Chen, H., Wang, Z., Li, Y.: Adamv-moe: Adaptive multi-task vision mixture-of-experts. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 17346–17357 (2023)
10. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
11. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
12. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
13. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Jain, S., et al.: Perceptionlm: Open-access data and models for detailed visual understanding. arXiv preprint arXiv:2504.13180 (2025)
14. Clark, C., Zhang, J., Ma, Z., Park, J.S., Tripathi, R., Lee, S., Salehi, M., Ren, J., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28652–28668 (2026)
15. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
16. Csordás, R., Irie, K., Schmidhuber, J.: Approximating two-layer feedforward networks for efficient transformers. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 674–692 (2023)
17. Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1280–1297 (2024)
18. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision

transformers to 22 billion parameters. In: International conference on machine learning. pp. 7480–7512. PMLR (2023)
19. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
20. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al.: Maskclip: Masked self-distillation advances contrastive language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10995–11005 (2023)
21. Dong, Y., Noci, L., Khodak, M., Li, M.: Attention retrieves, mlp memorizes: Disentangling trainable components in the transformer. arXiv e-prints pp. arXiv–2506 (2025)
22. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
23. Fini, E., Shukor, M., Li, X., Dufter, P., Klein, M., Haldimann, D., Aitharaju, S., da Costa, V.G.T., Béthune, L., Gan, Z., Toshev, A.T., Eichner, M., Nabi, M., Yang, Y., Susskind, J.M., El-Nouby, A.: Multimodal autoregressive pre-training of large vision encoders (2024), https://arxiv.org/abs/2411.14402
24. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis (2025), https://arxiv.org/abs/2405.21075
25. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
26. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
27. Gu, J., Stevens, S., Campolongo, E.G., Thompson, M.J., Zhang, N., Wu, J., Kopanev, A., Mai, Z., White, A.E., Balhoff, J., et al.: Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. arXiv preprint arXiv:2505.23883 (2025)
28. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
29. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
30. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15262–15271 (2021)
31. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
32. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)

33. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
34. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the em algorithm. Neural computation **6**(2), 181–214 (1994)
35. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
36. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
37. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
38. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
39. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
40. Li, J., Wang, X., Zhu, S., Kuo, C.W., Xu, L., Chen, F., Jain, J., Shi, H., Wen, L.: Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. Advances in Neural Information Processing Systems **37**, 131224–131246 (2024)
41. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Lou, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22195–22206 (2024)
42. Lin, B., Tang, Z., Ye, Y., Huang, J., Zhang, J., Pang, Y., Jin, P., Ning, M., Luo, J., Yuan, L.: Moe-llava: Mixture of experts for large vision-language models. IEEE Transactions on Multimedia (2026)
43. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
44. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
45. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
46. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
47. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: Advances in Neural Information Processing Systems (2023)
48. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 2263–2279. Association for Computational Linguistics (2022)
49. Mathew, M., Bagal, V., Tito, R.P., Karatzas, D., Valveny, E., Jawahar, C.V.: Infographicvqa (2021), `https://arxiv.org/abs/2104.12756`

50. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2199–2208 (2021)
51. Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., Houlsby, N.: Multimodal contrastive learning with limoe: the language-image mixture of experts. Advances in Neural Information Processing Systems **35**, 9564–9576 (2022)
52. Nguyen, H., Ho, N., Rinaldo, A.: Sigmoid gating is more sample efficient than softmax gating in mixture of experts. Advances in Neural Information Processing Systems **37**, 118357–118388 (2024)
53. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
54. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
56. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
57. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)
58. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
59. Shen, L., Chen, G., Shao, R., Guan, W., Nie, L.: MoME: Mixture of multimodal experts for generalist multimodal large language models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), `https://openreview.net/forum?id=Xskl7Da34U`
60. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European conference on computer vision. pp. 742–758. Springer (2020)
61. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., et al.: Dinov3 (2025), `https://arxiv.org/abs/2508.10104`
62. Singh, A., Pang, G., et al.: Textvqa: Towards reading and reasoning in visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
63. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
64. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2024)
65. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al.: Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
66. Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., et al.: Kimi k2: Open agentic intelligence. arXiv preprint arXiv:2507.20534 (2025)

67. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
68. Tillet, P., Kung, H.T., Cox, D.: Triton: an intermediate language and compiler for tiled neural network computations. In: Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages. pp. 10–19 (2019)
69. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
70. Wan, B., Tschannen, M., Xian, Y., Pavetic, F., Alabdulmohsin, I.M., Wang, X., Susano Pinto, A., Steiner, A., Beyer, L., Zhai, X.: Locca: Visual pretraining with location-aware captioners. Advances in Neural Information Processing Systems **37**, 116355–116387 (2024)
71. Wang, L., Gao, H., Zhao, C., Sun, X., Dai, D.: Auxiliary-loss-free load balancing strategy for mixture-of-experts. arXiv preprint arXiv:2408.15664 (2024)
72. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
73. Wang, X., Chen, C., Yang, Y., Chen, H.Y., Zhang, B., Pal, A., Zhu, X., Du, X.: Clip-up: A simple and efficient mixture-of-experts clip training recipe with sparse upcycling. arXiv preprint arXiv:2502.00965 (2025)
74. Wu, J., Hu, X., Wang, Y., Pang, B., Soricut, R.: 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. pp. 14205–14215 (2024)
75. Wu, L., Liu, M., Chen, Y., Chen, D., Dai, X., Yuan, L.: Residual mixture of experts. arXiv preprint arXiv:2204.09636 (2022)
76. Wu, Y., Du, J., Yan, K., Ding, S., Li, X.: ToVE: Efficient vision-language learning via knowledge transfer from vision experts. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=EMMnAd3apQ`
77. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023)
78. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
79. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
80. Ye, H., Yang, C.H.H., Goel, A., Huang, W., Zhu, L., Su, Y., Lin, S., Cheng, A.C., Wan, Z., Tian, J., et al.: Omnivinci: Enhancing architecture and data for omni-modal understanding llm. arXiv preprint arXiv:2510.15870 (2025)
81. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the association for computational linguistics **2**, 67–78 (2014)
82. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)

83. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: Merlot: Multimodal neural script knowledge models. Advances in neural information processing systems **34**, 23634–23651 (2021)
84. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
85. Zhang, J., Qu, X., Zhu, T., Cheng, Y.: Clip-moe: Towards building mixture of experts for clip with diversified multiplet upcycling. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 5406–5419 (2025)