

Scaling Laws for Native Multimodal Models

Mustafa Shukor² Enrico Fini¹ Victor Guilherme Turrissi da Costa¹ Matthieu Cord²
 Joshua Susskind¹ Alaaeldin El-Nouby¹
¹Apple ²Sorbonne University

Abstract

Building general-purpose models that can effectively perceive the world through multimodal signals has been a long-standing goal. Current approaches involve integrating separately pre-trained components, such as connecting vision encoders to LLMs and continuing multimodal training. While such approaches exhibit remarkable sample efficiency, it remains an open question whether such late-fusion architectures are inherently superior. In this work, we revisit the architectural design of native multimodal models (NMMs)—those trained from the ground up on all modalities—and conduct an extensive scaling laws study, spanning 457 trained models with different architectures and training mixtures. Our investigation reveals no inherent advantage to late-fusion architectures over early-fusion ones, which do not rely on image encoders or tokenizers. On the contrary, early-fusion exhibits stronger performance at lower parameter counts, is more efficient to train, and is easier to deploy. Motivated by the strong performance of the early-fusion architectures, we show that incorporating Mixture of Experts (MoEs) allows models to learn modality-specific weights, significantly benefiting performance.

1. Introduction

Multimodality provides a rich signal for perceiving and understanding the world. Advances in vision [23, 52, 55, 78] and language models [3, 19, 65] have enabled the development of powerful multimodal models that understand language, images, and audio. A common approach involves grafting separately pre-trained unimodal models, such as connecting a vision encoder to the input layer of an LLM [6, 9, 35, 43, 61, 71, 76].

Although this seems like a convenient approach, it remains an open question whether such late-fusion strategies are inherently optimal for understanding multimodal signals. Moreover, with abundant multimodal data available, initializing from unimodal pre-training is potentially detrimental, as it may introduce biases that prevent the model

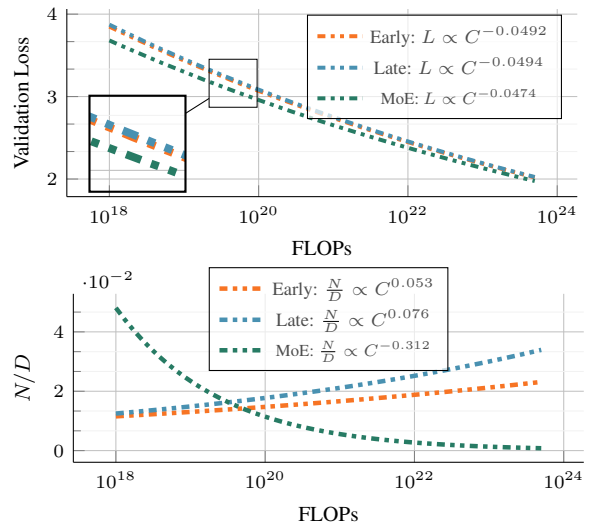


Figure 1. Scaling properties of Native Multimodal Models. Based on the scaling laws study in § 3.1, we observe: (1) early and late fusion models provide similar validation loss L when trained with the same compute budget C (FLOPs); (2) This performance is achieved via a different trade-off between parameters N and number of training tokens D , where early-fusion models requires fewer parameters. (3) Sparse early-fusion models achieve lower loss and require more training tokens for a given FLOP budget.

from fully leveraging cross-modality co-dependencies. An additional challenge is scaling such systems; each component (e.g., vision encoder, LLM) has its own set of hyperparameters, pre-training data mixtures, and scaling properties with respect to the amount of data and compute applied. A more flexible architecture might allow the model to dynamically allocate its capacity across modalities, simplifying scaling efforts.

In this work, we focus on the scaling properties of native multimodal models trained from the ground up on multimodal data. We first investigate whether the commonly adopted late-fusion architectures hold an intrinsic advantage by comparing them to early-fusion models, which process raw multimodal inputs without relying on dedicated vision encoders. We conduct scaling experiments on early and late fusion architectures, deriving scaling laws to pre-

dict their performance and compute-optimal configurations. Our findings indicate that late fusion offers no inherent advantage when trained from scratch. Instead, early-fusion models are more efficient and are easier to scale. Furthermore, we observe that native multimodal models follow scaling laws similar to those of LLMs [26], albeit with slight variations in scaling coefficients across modalities and datasets. Our results suggest that model parameters and training tokens should be scaled roughly equally for optimal performance. Moreover, we find that different multimodal training mixtures exhibit similar overall trends, indicating that our findings are likely to generalize to a broader range of settings.

While our findings favor early fusion, multimodal data is inherently heterogeneous, suggesting that some degree of parameter specialization may still offer benefits. To investigate this, we explore leveraging Mixture of Experts (MoEs) [59], a technique that enables the model to dynamically allocate specialized parameters across modalities in a symmetric and parallel manner, in contrast to late-fusion models, which are asymmetric and process data sequentially. Training native multimodal models with MoEs results in significantly improved performance and therefore, faster convergence. Our scaling laws for MoEs suggest that scaling number of training tokens is more important than the number of active parameters. This unbalanced scaling is different from what is observed for dense models, due to the higher number of total parameters for sparse models. In addition, Our analysis reveals that experts tend to specialize in different modalities, with this specialization being particularly prominent in the early and last layers.

1.1. Summary of our findings

Our findings can be summarized as follows:

Native Early and Late fusion perform on par: Early fusion models trained from scratch perform on par with their late-fusion counterparts, with a slight advantage to early-fusion models for low compute budgets (Figure 8). Furthermore, our scaling laws study indicates that the compute-optimal models for early and late fusion perform similarly as the compute budget increases (Figure 1 Top).

NMMs scale similarly to LLMs: The scaling laws of native multimodal models follow similar laws as text-only LLMs with slightly varying scaling exponents depending on the target data type and training mixture (Table 2).

Late-fusion requires more parameters: Compute-optimal late-fusion models require a higher parameters-to-data ratio when compared to early-fusion (Figure 1 bottom).

Sparsity significantly benefits early-fusion NMMs: Sparse NMMs exhibit significant improvements compared to their dense counterparts at the same inference cost (Figure 10). Furthermore, they implicitly learn modality-specific weights when trained with sparsity (Figure 25). In

Expression	Definition
N	Number of parameters in the multimodal decoder. For MoEs this refers to the active parameters only.
D	Total number of multimodal tokens.
N_v	Number of vision-only tokens.
D_v	Number of parameters in the vision-specific encoder. Only exists in late-fusion architectures.
C	Total number of FLOPs, estimated as $C = 6ND$ for early-fusion and $C = 6(N_v D_v + ND)$ for late-fusion.
L	Validation loss measured as the average over interleaved image-text, image-caption, and text-only data mixtures.

Table 1. Definitions of the expressions used throughout the paper.

addition, compute-optimal models rely more on scaling the number of training tokens than the number of active parameters as the compute-budget grows (Figure 1 Right).

Modality-agnostic routing beats Modality-aware routing for Sparse NMMs: Training sparse mixture of experts with modality-agnostic routing consistently outperforms models with modality-aware routing (Figure 11).

2. Preliminaries

2.1. Definitions

Native Multimodal Models (NMMs): Models that are trained from scratch on all modalities simultaneously without relying on pre-trained LLMs or vision encoders. Our focus is on the representative image and text modalities, where the model processes both text and images as input and generates text as output.

Early fusion: Enabling multimodal interaction from the beginning, using almost no modality-specific parameters (*e.g.*, except a linear layer to patchify images). Using a single transformer model, this approach processes raw multimodal input—tokenized text and continuous image patches—with no image discretization. In this paper, we refer to the main transformer as the decoder.

Late fusion: Delaying the multimodal interaction to deeper layers, typically after separate unimodal components has processed that process each modality independently (*e.g.*, a vision encoder connected to a decoder).

Modality-agnostic routing: In sparse mixture-of-experts, modality-agnostic routing refers to relying on a learned router module that is trained jointly with the model.

Modality-aware routing: Routing based on pre-defined rules such as routing based on the modality type (*e.g.*, vision-tokens, token-tokens).

2.2. Scaling Laws

We aim to understand the scaling properties of NMMs and how different architectural choices influence trade-offs. To this end, we analyze our models within the scaling laws framework proposed by Hoffmann et al. [26], Kaplan et al. [31]. We compute FLOPs based on the total number of parameters, using the approximation $C = 6ND$, as adopted in prior work [2, 26]. However, we modify this estimation to suit our setup: for late-fusion models, FLOPs is computed

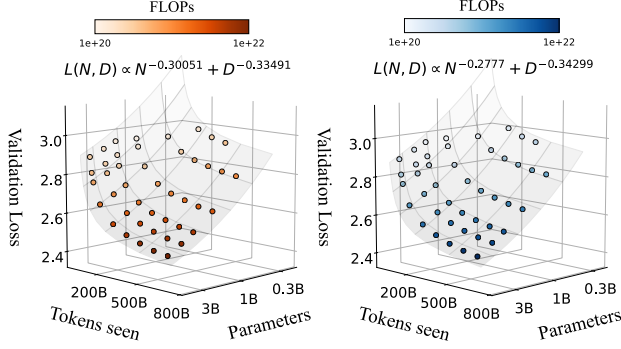


Figure 2. Scaling laws for early-fusion and late-fusion native multimodal models. Each point represents a model (300M to 3B parameters) trained on varying number of tokens (250M to 400B). We report the average cross-entropy loss on the validation sets of interleaved (Obelics), Image-caption (HQITP), and text-only data (DCLM).

as $6(N_v D_v + ND)$. We consider a setup where, given a compute budget C , our goal is to predict the model’s final performance, as well as determine the optimal number of parameters or number of training tokens. Consistent with prior studies on LLM scaling [26], we assume a power-law relationship between the final model loss and both model size (N) and training tokens (D):

$$L = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}. \quad (1)$$

Here, E represents the lowest achievable loss on the dataset, while $\frac{A}{N^\alpha}$ captures the effect of increasing the number of parameters, where a larger model leads to lower loss, with the rate of improvement governed by α . Similarly, $\frac{B}{D^\beta}$ accounts for the benefits of a higher number of tokens, with β determining the rate of improvement. Additionally, we assume a linear relationship between compute budget (FLOPs) and both N and D ($C \propto ND$). This further leads to power-law relationships detailed in Appendix C.7.

2.3. Experimental setup

Our models are based on the autoregressive transformer architecture [69] with SwiGLU FFNs [58] and QK-Norm [17] following Li et al. [39]. In early-fusion models, image patches are linearly projected to match the text token dimension, while late-fusion follows the CLIP architecture [55]. We adopt causal attention for text tokens and bidirectional attention for image tokens, we found this to work better. Training is conducted on a mixture of public and private multimodal datasets, including DCLM [39], Obelics [34], DFN [21], COYO [11], and a private collection of High-Quality Image-Text Pairs (HQITP). Images are resized to 224×224 resolution with a 14×14 patch size. We use a context length of 1k for the multimodal sequences. For training efficiency, we train our models with bfloat16, Fully Sharded Data Parallel (FSDP) [80], activation checkpointing, and gradient accumulation. We also use se-

		$L = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$	$N \propto C^a$	$D \propto C^b$	$L \propto C^c$	$D \propto N^d$		
Model	Data	E	α	β	a	b	c	d
GPT3 [10]	Text	–	–	–	–	–	-0.048	
Chinchilla [26]	Text	1.693	0.339	0.285	0.46	0.54	–	
NMM (early-fusion)	Text	2.222	0.3084	0.3375	0.5246	0.4774	-0.0420	0.9085
	Image-Caption	1.569	0.3111	0.3386	0.5203	0.4785	-0.0610	0.9187
	Interleaved	1.966	0.2971	0.338	0.5315	0.4680	-0.0459	0.8791
	AVG	1.904	0.301	0.335	0.5262	0.473	-0.0492	0.8987
NMM (late-fusion)	AVG	1.891	0.2903	0.3383	0.6358	0.4619	-0.0494	0.6732
Sparse NMM (early-fusion)	AVG	2.158	0.710	0.372	0.361	0.656	-0.047	1.797

Table 2. Scaling laws for native multimodal models. We report the scaling laws results for early and late fusion models. We fit the scaling laws for different target data types as well as their average loss (AVG).

quence packing for the image captioning dataset to reduce the amount of padded tokens. Similar to previous works [2, 5, 26], we evaluate performance on held-out subsets of interleaved (Obelics), Image-caption (HQITP), and text-only data (DCLM). Further implementation details are provided in Appendix A.

3. Scaling native multimodal models

In this section, we present a scaling laws study of native multimodal models, examining various architectural choices § 3.1, exploring different data mixtures § 3.2, analyzing the practical trade-offs between late and early fusion NMMs, and comparing the performance of native pre-training and continual pre-training of NMMs § 3.3.

Setup. We train models ranging from 0.3B to 4B active parameters, scaling the width while keeping the depth constant. For smaller training token budgets, we reduce the warm-up phase to 1K steps while maintaining 5K steps for larger budgets. Following Hägele et al. [25], models are trained with a constant learning rate, followed by a cool-down phase using an inverse square root scheduler. The cool-down phase spans 20% of the total steps spent at the constant learning rate. To estimate the scaling coefficients in Eq 1, we apply the L-BFGS algorithm [51] and Huber loss [28] (with $\delta = 10^{-3}$), performing a grid search over initialization ranges.

3.1. Scaling laws of NMMs

Scaling laws for early-fusion and late-fusion models. Figure 2 (left) presents the final loss averaged across interleaved, image-caption, and text datasets for early-fusion NMMs. The lowest-loss frontier follows a power law as a function of FLOPs. Fitting the power law yields the expression $L \propto C^{-0.049}$, indicating the rate of improvement with increasing compute. When analyzing the scaling laws per data type (e.g., image-caption, interleaved, text), we observe that the exponent varies (Table 2). For instance, the model achieves a higher rate of improvement for image-

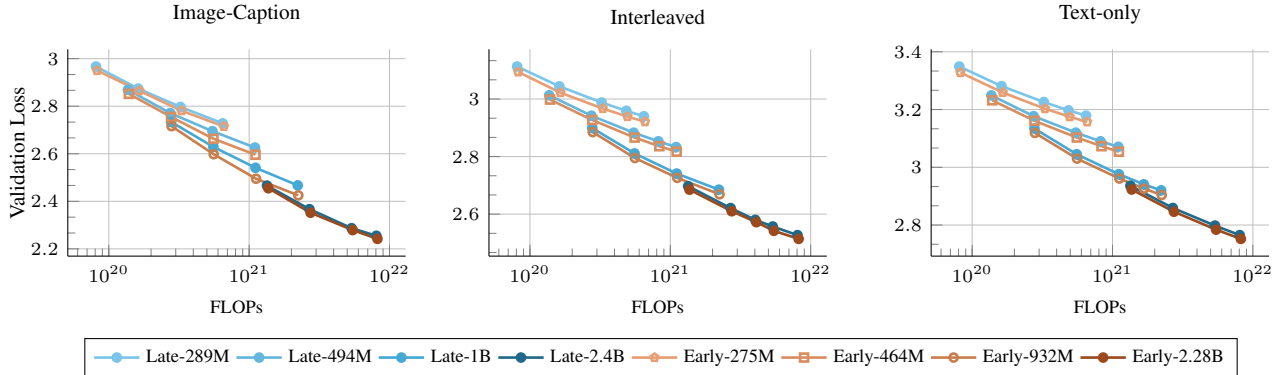


Figure 3. Early vs late fusion: scaling training FLOPs. We compare early and late fusion models when scaling both the number of model parameters and the number of training tokens. Overall, early fusion shows a slight advantage, especially at smaller model sizes, and the gap decreases when scaling the number of parameters N .

caption data ($L \propto C^{-0.061}$) when compared to interleaved documents ($L \propto C^{-0.046}$).

To model the loss as a function of the number of training tokens D and model parameters N , we fit the parametric function in Eq 1, obtaining scaling exponents $\alpha = 0.301$ and $\beta = 0.335$. These describe the rates of improvement when scaling the number of model parameters and training tokens, respectively. Assuming a linear relationship between compute, N , and D (i.e., $C \propto ND$), we derive the law relating model parameters to the compute budget (see Appendix C for details). Specifically, for a given compute budget C , we compute the corresponding model size N at logarithmically spaced D values and determine N_{opt} , the parameter count that minimizes loss. Repeating this across different FLOPs values produces a dataset of (C, N_{opt}) , to which we fit a power law predicting the compute-optimal model size as a function of compute: $N^* \propto C^{0.526}$.

Similarly, we fit power laws to estimate the compute-optimal training dataset size as a function of compute and model size:

$$D_{opt} \propto C^{0.473}, \quad D_{opt} \propto N^{0.899}.$$

These relationships allow practitioners to determine the optimal model and dataset size given a fixed compute budget. When analyzing by data type, we find that interleaved data benefits more from larger models ($a = 0.532$) compared to image-caption data ($a = 0.520$), whereas the opposite trend holds for training tokens.

We conduct a similar study on late-fusion models in Figure 2 (right) and observe comparable scaling behaviors. In particular, the loss scaling exponent ($c = -0.0494$) is nearly identical to that of early fusion ($c = -0.0492$). This trend is evident in Figure 3, where early fusion outperforms late fusion at smaller model scales, while both architectures converge to similar performance at larger model sizes. We also observe similar trends when varying late-fusion con-

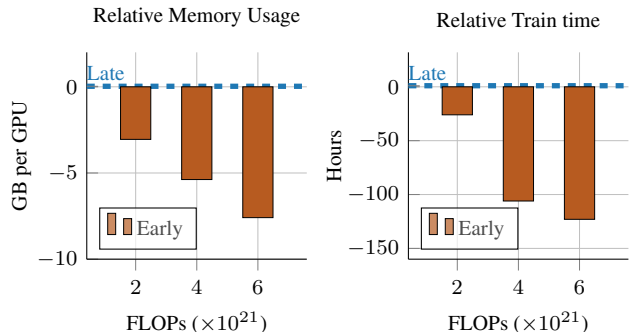


Figure 4. Early vs late: pretraining efficiency. Early-fusion is faster to train and consumes less memory. Models are trained on 16 H100 GPUs for 160k steps (300B tokens).

figurations, such as using a smaller vision encoder with a larger text decoder Appendix B.

Scaling laws of NMMs vs LLMs. Upon comparing the scaling law coefficients of our NMMs to those reported for text-only LLMs (e.g., GPT-3, Chinchilla), we find them to be within similar ranges. In particular, for predicting the loss as a function of compute, GPT-3 [10] follows $L \propto C^{-0.048}$, while our models follow $L \propto C^{-0.049}$, suggesting that the performance of NMMs adheres to similar scaling laws as LLMs. Similarly, our estimates of the α and β parameters in Eq 1 ($\alpha = 0.301$, $\beta = 0.335$) closely match those reported by Hoffmann et al. [26] ($\alpha = 0.339$, $\beta = 0.285$). Likewise, our computed values of $a = 0.526$ and $b = 0.473$ align closely with $a = 0.46$ and $b = 0.54$ from [26], reinforcing the idea that, for native multimodal models, the number of training tokens and model parameters should be scaled proportionally. However, since the gap between a and b is smaller than in LLMs, this principle holds even more strongly for NMMs. Additionally, as $a = 0.526$ is greater than $b = 0.473$ in our case, the optimal model size for NMMs is larger than that of LLMs,

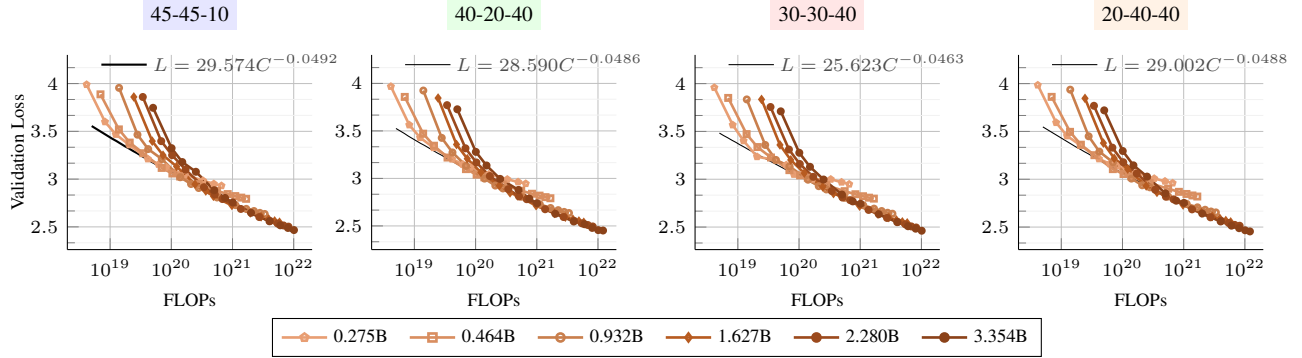


Figure 5. Scaling laws with different training mixtures. Early-fusion models follow similar scaling trends when changing the pretraining mixtures. However, increasing the image captions leads to a higher scaling exponent norm (see Table 3).

	C-I-T (%)	I/T ratio	E	α	β	a	b	d	c
1	45-45-10	1.19	1.906	0.301	0.335	0.527	0.474	0.901	-0.0492
2	40-20-40	0.65	1.965	0.328	0.348	0.518	0.486	0.937	-0.0486
3	30-30-40	0.59	1.847	0.253	0.338	0.572	0.428	0.748	-0.0463
4	20-40-40	0.49	1.836	0.259	0.354	0.582	0.423	0.726	-0.0488

Table 3. Scaling laws for different training mixtures. Early-fusion models. C-I-T refer to image-caption, interleaved and text while the optimal number of training tokens is lower, given a fixed compute budget.

Compute-optimal trade-offs for early vs. late fusion NMMs. While late- and early-fusion models reduce loss at similar rates with increasing FLOPs, we observe distinct trade-offs in their compute-optimal models. Specifically, N_{opt} is larger for late-fusion models, whereas D_{opt} is larger for early-fusion models. This indicates that, given a fixed compute budget, late-fusion models require a higher number of parameters, while early-fusion models benefit more from a higher number of training tokens. This trend is also reflected in the lower $\frac{N_{opt}}{D_{opt}} \propto C^{0.053}$ for early fusion compared to $\frac{N_{opt}}{D_{opt}} \propto C^{0.076}$ for late fusion. As shown in Figure 1 (bottom), when scaling FLOPs, the number of parameters of early fusion models becomes significantly lower, which is crucial for reducing inference costs and, consequently, lowering serving costs after deployment.

Early-fusion is more efficient to train. We compare the training efficiency of late- and early-fusion architectures. As shown in Figure 4, early-fusion models consume less memory and train faster under the same compute budget. This advantage becomes even more pronounced as compute increases, highlighting the superior training efficiency of early fusion while maintaining comparable performance to late fusion at scale. Notably, for the same FLOPs, late-fusion models have a higher parameter count and higher effective depth (*i.e.*, additional vision encoder layers alongside decoder layers) compared to early-fusion models.

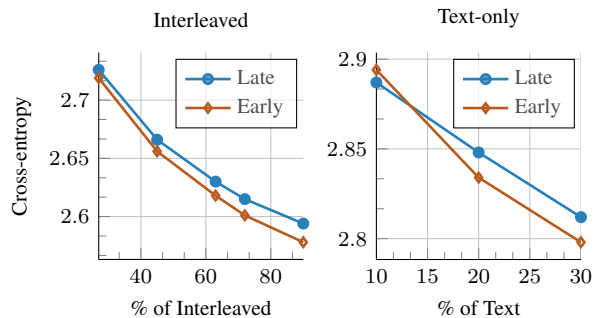


Figure 7. Early vs late fusion: changing the training mixture. We vary the training mixtures and plot the final training loss. Early fusion models attain a favorable performance when increasing the proportion of interleaved documents and text-only data.

3.2. Scaling laws for different data mixtures

We investigate how variations in the training mixture affect the scaling laws of native multimodal models. To this end, we study four different mixtures that reflect common community practices [34, 41, 46, 79], with Image Caption-Interleaved-Text ratios of 45-45-10 (our default setup), 30-30-40, 40-20-40, and 20-40-40. For each mixture, we conduct a separate scaling study by training 76 different models, following our setup in § 3.1. Overall, Figure 5 shows that different mixtures follow similar scaling trends; however, the scaling coefficients vary depending on the mixture (Table 3). Interestingly, increasing the proportion of image-caption data (mixtures 1 and 2) leads to lower a and higher b , whereas increasing the ratio of interleaved and text data (mixtures 3 and 4) have the opposite effect. Notably, image-caption data contains more image tokens than text tokens; therefore, increasing its proportion results in more image tokens, while increasing interleaved and text data increases text token counts. This suggests that, when image tokens are prevalent, training for longer decreases the loss faster than increasing the model size. We also found that for a fixed model size, increasing text-only and interleaved data ratio is in favor of early-fusion Figure 7.

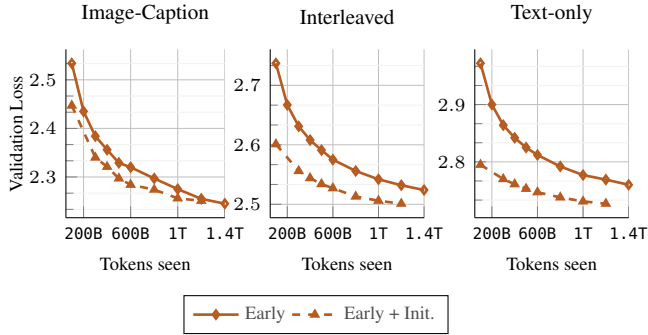


Figure 8. Early native vs initializing from LLMs: initializing from pre-trained models and scaling training tokens. We compare training with and without initializing from DCLM-1B.

3.3. Native multimodal pre-training vs. continual training of LLMs

In this section, we compare training natively from scratch to continual training after initializing from a pre-trained LLM. We initialize the model from DCLM-1B [21] that is trained on more than 2T tokens. Figure 8 shows that native multimodal models can close the gap with initialized models when trained for longer. Specifically, on image captioning data, the model requires fewer than 100B multimodal tokens to reach comparable performance. However, on interleaved and text data, the model may need longer training—up to 1T tokens. Considering the cost of pre-training, these results suggest that training natively could be a more efficient approach for achieving the same performance on multimodal benchmarks.

4. Towards multimodal specialization

Previously, we demonstrated that early-fusion models achieve performance on par with late-fusion models under a fixed compute budget. However, multimodal data is inherently heterogeneous, and training a unified model to fit such diverse distributions may be suboptimal. Here, we argue for multimodal specialization within a unified architecture. Ideally, the model should implicitly adapt to each modality, for instance, by learning modality-specific weights or specialized experts. Mixture of Experts is a strong candidate for this approach, having demonstrated effectiveness in LLMs. In this section, we highlight the advantages of sparse early-fusion models over their dense counterparts.

Setup. Our sparse models are based on the dropless-MoE implementation of Gale et al. [24], which eliminates token dropping during training caused by expert capacity constraints. We employ a top- k expert-choice routing mechanism, where each token selects its top- k experts among the E available experts. Specifically, we set $k = 1$ and $E = 8$, as we find this configuration to work effectively. Additionally, we incorporate an auxiliary load-balancing loss [59] with a weight of 0.01 to ensure a balanced expert utilization.

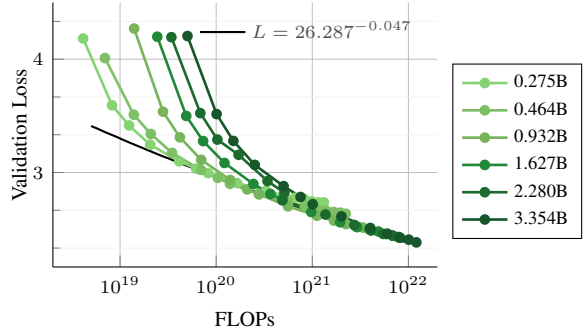


Figure 9. Scaling laws for sparse early-fusion NMMs. We report the final validation loss averaged across interleaved, image-captions and text data.

Following Abnar et al. [2], we compute training FLOPs as $6ND$, where N represents the number of active parameters.

4.1. Sparse vs dense NMMs when scaling FLOPs

We compare sparse MoE models to their dense counterparts by training models with different numbers of active parameters and varying amounts of training tokens. Figure 10 shows that, under the same inference cost (or number of active parameters), MoEs significantly outperform dense models. Interestingly, this performance gap is more pronounced for smaller model sizes. This suggests that MoEs enable models to handle heterogeneous data more effectively and specialize in different modalities. However, as dense models become sufficiently large, the gap between the two architectures gradually closes.

4.2. Scaling laws for sparse early-fusion models

We train different models (ranging from 300M to 3.4B active parameters) on varying amounts of tokens (ranging from 250M to 600B) and report the final loss in Figure 9. We fit a power law to the convex hull of the lowest loss as a function of compute (FLOPs). Interestingly, the exponent (-0.048) is close to that of dense NMMs (-0.049), indicating that both architectures scale similarly. However, the multiplicative constant is smaller for MoEs (27.086) compared to dense models (29.574), revealing lower loss. Additionally, MoEs require longer training to reach saturation compared to dense models (Appendix C for more details). We also predict the coefficients of Eq 1 by considering N as the number of active parameters. Table 2 shows significantly higher α compared to dense models. Interestingly, b is significantly higher than a , revealing that the training tokens should be scaled at a higher rate than the number of parameters when training sparse NMMs. We also experiment with a scaling law that takes into account the sparsity [2] and reached similar conclusions Appendix C.7.

4.3. Modality-aware vs. Modality-agnostic routing

Another alternative to MoEs is modality-aware routing, where multimodal tokens are assigned to experts based on

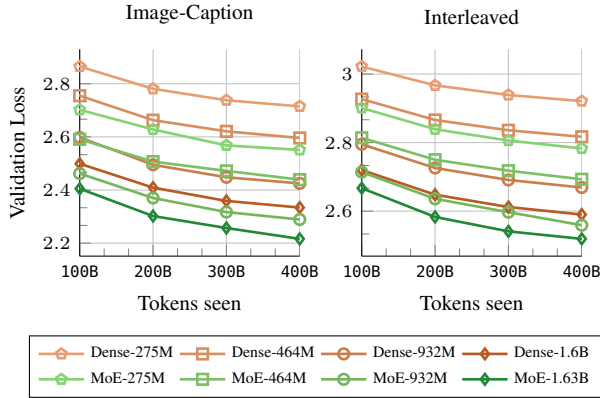


Figure 10. MoE vs Dense: scaling training FLOPs. We compare MoE and dense early-fusion models when scaling both the amount of training tokens and model sizes. MoEs beat dense models when matching the number of active parameters.

their modalities, similar to previous works [7, 73]. We train models with distinct image and text experts in the form of FFNs, where image tokens are processed only by the image FFN and text tokens only by the text FFN. Compared to modality-aware routing, MoEs exhibit significantly better performance on both image-caption and interleaved data as presented in Figure 11.

4.4. Emergence of expert specialization and sharing

We investigate multimodal specialization in MoE architectures. In Figure 13, we visualize the normalized number of text and image tokens assigned to each expert across layers. To quantify this specialization, we compute a specialization score, defined as the average, across all experts within a layer, of $1 - H(p)$, where H is the binary entropy of each expert’s text/image token distribution. We plot this specialization score in Figure 12. Higher specialization scores indicate a tendency for experts to focus on either text or image tokens, while lower scores indicate a shared behavior. These visualizations provide clear evidence of modality-specific experts, particularly in the early layers. Furthermore, the specialization score decreases as the number of layers increases, before rising again in the last layers. This suggests that early and final layers exhibit higher modality specialization compared to mid-layers. This behavior is intuitive, as middle layers are expected to hold higher-level features that may generalize across modalities, and consistent with findings in [?] that shows increasing alignment between modalities across layers. The emergence of both expert specialization and cross-modality sharing in our modality-agnostic MoE, suggests it may be a preferable approach compared to modality-aware sparsity. All data displayed here is from an early-fusion MoE model with 1B active parameters trained for 300B tokens.

	Accuracy					CIDEr		
	AVG	VQA _{v2}	TextVQA	OKVQA	GQA	VizWiz	COCO	TextCaps
Late-fusion	46.8	69.4	25.8	50.1	65.8	22.8	70.7	50.9
Early-fusion	47.6	69.3	28.1	52.1	65.4	23.2	72.0	53.8
Early-MoEs	48.2	69.8	30.0	52.1	65.4	23.6	69.6	55.7

Table 4. Supervised finetuning on the LLaVA mixture. All models are native at 1.5B scale and pre-trained on 300B tokens.

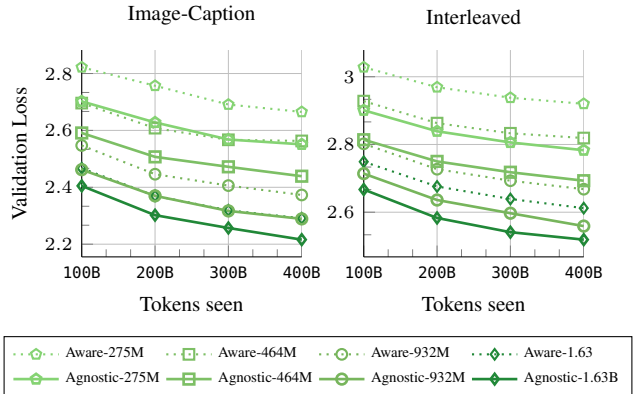


Figure 11. Modality-aware vs modality agnostic routing for sparse NMMs. We compare modality-agnostic routing with modality-aware routing when scaling both the amount of training tokens and model sizes.

5. Evaluation on downstream tasks with SFT

Following previous work on scaling laws, we primarily rely on validation losses. However, we generally find that this evaluation correlates well with performance on downstream tasks. To validate this, we conduct a multimodal instruction tuning stage (SFT) on the LLaVA mixture [43] and report accuracy and CIDEr scores across several VQA and captioning tasks. Table 4 confirms the ranking of different model configurations. Specifically, early fusion outperforms late fusion, and MoEs outperform dense models. However, since the models are relatively small (1.5B scale), trained from scratch, and fine-tuned on a small dataset, the overall scores are lower than the current state of the art. Further implementation details can be found in Appendix A.

6. Related work

Large multimodal models. A long-standing research goal has been to develop models capable of perceiving the world through multiple modalities, akin to human sensory experience. Recent progress in vision and language processing has shifted the research focus from smaller, task-specific models toward large, generalist models that can handle diverse inputs [29, 65]. Crucially, pre-trained vision and language backbones often require surprisingly little adaptation to enable effective cross-modal communication [32, 47, 61, 66, 67]. Simply integrating a vision encoder with either an encoder-decoder architecture [45, 48, 62, 70]

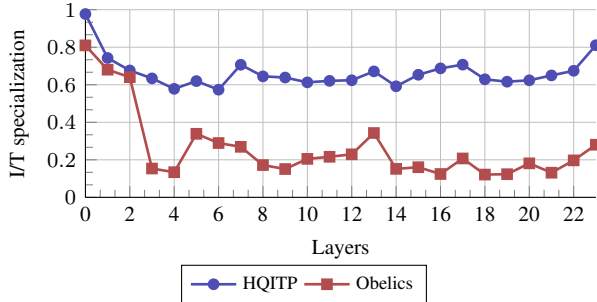


Figure 12. MoE specialization score. Entropy-based image/text specialization score (as described in § 4.4) across layers for two data sources: HQITP and Obelics. HQITP has a more imbalanced image-to-text token distribution, resulting in generally higher specialization. Despite this difference, both data sources exhibit a similar trend: the specialization score decreases in the early layers before increasing again in the final layers.

or a decoder-only LLM has yielded highly capable multimodal systems [1, 6, 9, 13, 16, 35, 43, 49, 71, 76, 81]. This late-fusion approach, where modalities are processed separately before being combined, is now well-understood, with established best practices for training effective models [34, 41, 46, 79]. In contrast, early-fusion models [8, 18, 64], which combine modalities at an earlier stage, remain relatively unexplored, with only a limited number of publicly released models [8, 18]. Unlike [18, 64], our models utilize only a single linear layer and rely exclusively on a next-token prediction loss. Furthermore, we train our models from scratch on all modalities without image tokenization.

Native Multimodal Models. We define native multimodal models as those trained from scratch on all modalities simultaneously [65] rather than adapting LLMs to accommodate additional modalities. Due to the high cost of training such models, they remain relatively underexplored, with most relying on late-fusion architectures [27, 77]. Some multimodal models trained from scratch [4, 64, 74] relax this constraint by utilizing pre-trained image tokenizers such as [20, 68] to convert images into discrete tokens, integrating them into the text vocabulary. This approach enables models to understand and generate text and images, facilitating a more seamless multimodal learning process.

Scaling laws. Scaling law studies aim to predict how model performance scales with training compute. Early works [26, 31] found that LLM performance follows a power-law relationship with compute, enabling the compute-optimal estimation of the number of model parameters and training tokens at scale for a given budget. Similar research has extended these findings to sparse Mixture of Experts (MoE) models, considering factors such as sparsity, number of experts, and routing granularity [15, 33, 72]. Scaling laws have also been observed across various domains, including image models [23], video models [56], protein LLMs [14], and imitation learning [54]. However, few stud-

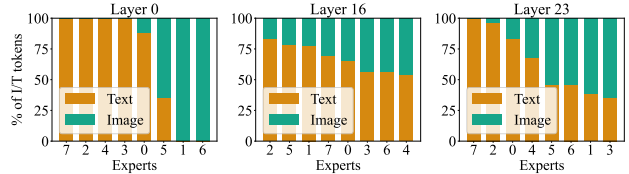


Figure 13. MoE specialization frequency. Percentage of text and image tokens routed to each expert on interleaved data from Obelics. Experts are ordered for better visualization. The first layer shows the highest amount of unimodal experts.

ies have investigated scaling laws for multimodal models. Notably, Aghajanyan et al. [5] examined multimodal models that tokenize modalities into discrete tokens and include multimodal generation. In contrast, we focus on studying early-fusion models that take raw multimodal inputs and are trained on interleaved multimodal data.

Mixture of experts (MoEs). MoEs [59] scale model capacity efficiently by sparsely activating parameters, enabling large models with reduced per-sample compute. While widely studied in LLMs [22, 30, 36, 37, 42, 63, 75, 82], MoEs remain underexplored in multimodal settings. Prior work has examined contrastive models [50], late-fusion LLMs [38, 40], and modality-specific experts [7, 12, 60]. We focus on analyzing MoEs in early-fusion multimodal models.

7. Limitations

Our study finds that scaling law coefficients are broadly consistent across training mixtures, though a broader exploration is needed to validate this observation. While validation loss scales predictably with compute, the extent to which this correlates with downstream performance remains unclear and warrants further investigation. The accuracy of scaling law predictions improves with higher FLOPs, but their extrapolation to extreme model sizes is still an open question (Appendix D for more details).

8. Conclusion

We explore various strategies for compute-optimal pretraining of native multimodal models. We found the NMMs follow similar scaling laws to those of LLMs. Contrary to common belief, we find no inherent advantage in adopting late-fusion architectures over early-fusion ones. While both architectures exhibit similar scaling properties, early-fusion models are more efficient to train and outperform late-fusion models at lower compute budgets. Furthermore, we show that sparse architectures encourage modality-specific specialization, leading to performance improvements while maintaining the same inference cost.

Acknowledgment

We thank Philipp Dufter, Samira Abnar, Xiujun Li, Zhe Gan, Alexander Toshev, Yinfei Yang, Dan Busbridge, and Jason Ramapuram for many fruitful discussions. We thank Denise Hui, and Samy Bengio for infra and compute support. Finally, we thank, Louis Béthune, Pierre Ablin, Marco Cuturi, and the MLR team at Apple for their support throughout the project.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadallah, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 8
- [2] Samira Abnar, Harshay Shah, Dan Busbridge, Alaaeldin Mohamed Elnouby Ali, Josh Susskind, and Vimal Thilak. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models. *arXiv preprint arXiv:2501.12370*, 2025. 2, 3, 6, 8
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [4] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022. 8
- [5] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023. 3, 8
- [6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 8
- [7] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlm0: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 7, 8
- [8] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Saĝnak Taşirlar. Introducing our multimodal models, 2023. 8
- [9] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarelli, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1, 8
- [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3, 4
- [11] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>, 2022. 3, 1
- [12] Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1110–1119, 2024. 8
- [13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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*, pages 24185–24198, 2024. 8
- [14] Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. *bioRxiv*, 2024. 8
- [15] Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pages 4057–4086. PMLR, 2022. 8
- [16] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 8
- [17] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 3
- [18] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024. 8
- [19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1
- [20] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 8
- [21] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 3, 6, 1

- [22] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. [8](#)
- [23] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrise da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders, 2024. [1](#), [8](#)
- [24] Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. Megablocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems*, 5:288–304, 2023. [6](#)
- [25] Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *arXiv preprint arXiv:2405.18392*, 2024. [3](#)
- [26] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030, 2022. [2](#), [3](#), [4](#), [8](#), [5](#)
- [27] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. [8](#)
- [28] Peter J. Huber. *Robust Estimation of a Location Parameter*, pages 492–518. Springer New York, New York, NY, 1992. [3](#)
- [29] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. [7](#)
- [30] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. [8](#)
- [31] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. [2](#), [8](#), [3](#)
- [32] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pages 17283–17300. PMLR, 2023. [7](#)
- [33] Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, et al. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*, 2024. [8](#), [6](#)
- [34] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. [3](#), [5](#), [8](#), [1](#)
- [35] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. [1](#), [8](#)
- [36] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. [8](#)
- [37] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pages 6265–6274. PMLR, 2021. [8](#)
- [38] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. [8](#)
- [39] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-1m: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024. [3](#), [1](#)
- [40] Bin Lin, Zhenyu Tang, Yang Ye, Jiayi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. [8](#)
- [41] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. [5](#), [8](#)
- [42] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. [8](#)
- [43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. [1](#), [7](#), [8](#)
- [44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [1](#)
- [45] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022. [7](#)
- [46] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xi-anzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2025. [5](#), [8](#), [1](#)

- [47] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*, 2023. 7
- [48] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023. 7
- [49] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Any-mal: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1314–1332, 2024. 8
- [50] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022. 8
- [51] Jorge Nocedal. Updating quasi newton matrices with limited storage. *Mathematics of Computation*, 35(151):951–958, 1980. 3
- [52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1
- [53] Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws. *arXiv preprint arXiv:2406.12907*, 2024. 3
- [54] Tim Pearce, Tabish Rashid, Dave Bignell, Raluca Georgescu, Sam Devlin, and Katja Hofmann. Scaling laws for pre-training agents and world models. *arXiv preprint arXiv:2411.04434*, 2024. 8
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [56] Jathushan Rajasegaran, Ilija Radosavovic, Rahul Ravishanker, Yossi Gandelsman, Christoph Feichtenhofer, and Jitendra Malik. An empirical study of autoregressive pre-training from videos. *arXiv preprint arXiv:2501.05453*, 2025. 8
- [57] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5571–5584, 2023. 1
- [58] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 3
- [59] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarsz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2, 6, 8
- [60] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 8
- [61] Mustafa Shukor, Corentin Dancette, and Matthieu Cord. ep-alm: Efficient perceptual augmentation of language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22056–22069, 2023. 1, 7
- [62] Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. Unival: Unified model for image, video, audio and language tasks. *Transactions on Machine Learning Research Journal*, 2023. 7
- [63] Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, et al. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*, 2024. 8
- [64] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 8
- [65] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 7, 8
- [66] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 7
- [67] Théophane Vallaey, Mustafa Shukor, Matthieu Cord, and Jakob Verbeek. Improved baselines for data-efficient perceptual augmentation of llms. *arXiv preprint arXiv:2403.13499*, 2024. 7
- [68] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 8
- [69] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3
- [70] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 7
- [71] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 8
- [72] Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, and Jingang Wang. Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*,

- pages 5583–5595, Miami, Florida, USA, 2024. Association for Computational Linguistics. 8, 6
- [73] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 7
- [74] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 8
- [75] Tianwen Wei, Bo Zhu, Liang Zhao, Cheng Cheng, Biye Li, Weiwei Lü, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Liang Zeng, et al. Skywork-moe: A deep dive into training techniques for mixture-of-experts language models. *arXiv preprint arXiv:2406.06563*, 2024. 8
- [76] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 1, 8
- [77] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 8
- [78] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1
- [79] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruvi Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mml. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 5, 8
- [80] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 3
- [81] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 8
- [82] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. Stmoe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. 8