
Scaling Laws for Native Multimodal Models

Mustafa Shukor*
Sorbonne University

Enrico Fini
Apple

Victor Guilherme Turrise da Costa
Apple

Matthieu Cord
Sorbonne University

Joshua Susskind
Apple

Alaaeldin El-Nouby
Apple

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. 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 for models that learn modality-specific weights, significantly enhancing performance.

1 Introduction

Multimodality provides a rich signal for perceiving and understanding the world. Advances in vision [Radford et al., 2021; Oquab et al., 2023; Zhai et al., 2023; Fini et al., 2024], audio [Huang et al., 2022; Elizalde et al., 2023; Chen et al., 2022; Hsu et al., 2021] and language models [Achiam et al., 2023; Team et al., 2023; Dubey et al., 2024] 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 [Laurençon et al., 2024b; Shukor et al., 2023a; Alayrac et al., 2022; Xue et al., 2024; Beyer et al., 2024; Wang et al., 2024a; Liu et al., 2024b; Zhang et al., 2023; Kong et al., 2024; Défossez et al., 2024].

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 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

*Work done during an internship at Apple.

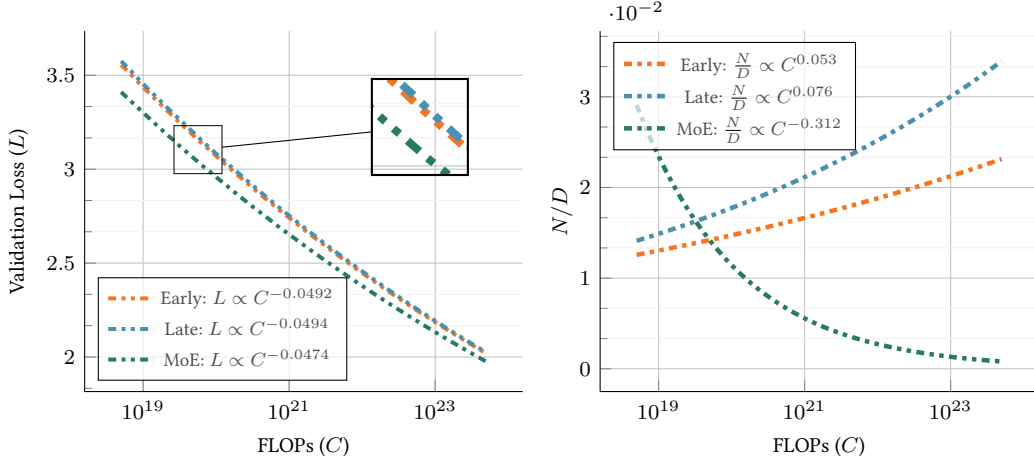


Figure 1: Scaling properties of Native Multimodal Models. Based on the scaling laws study in § 3.1, we observe: (1) early and late fusion models provide on par validation loss L when trained using the same compute budget C (in FLOPs); (2) This performance is achieved via a different trade-off between 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.

raw multimodal inputs without relying on dedicated vision encoders. We conduct scaling experiments on early and late fusion architectures, deriving scaling laws to predict 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 [Hoffmann et al., 2022], 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) [Shazeer et al., 2017], 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 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 (fig. 8). Furthermore, our scaling laws study indicates that the compute-optimal models for early and late fusion perform similarly as the compute budget increases (fig. 1 Left).

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 3).

Expression	Definition
N	Number of parameters in the multimodal decoder. For MoEs this refers to the active parameters.
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	Average validation loss on interleaved image-text, image-caption, and text-only data mixtures.

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

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

Sparsity significantly benefits early-fusion NMMs: Sparse NMMs exhibit significant improvements compared to their dense counterparts at the same inference cost (fig. 9). Furthermore, they implicitly learn modality-specific weights when trained with sparsity (fig. 23). In addition, compute-optimal models rely more on scaling the number of training tokens than the number of active parameters as the compute-budget grows (fig. 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 (fig. 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. 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 each modality independently (*e.g.*, a vision encoder connected to an LLM).

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 Kaplan et al. [2020]; Hoffmann et al. [2022]. We compute FLOPs based on the total number of parameters, using the approximation $C = 6ND$, as adopted in prior work [Hoffmann et al., 2022; Abnar et al., 2025]. However, we modify this estimation to suit our setup: for late-fusion models, FLOPs is computed 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 loss, as well as determine the optimal number of parameters

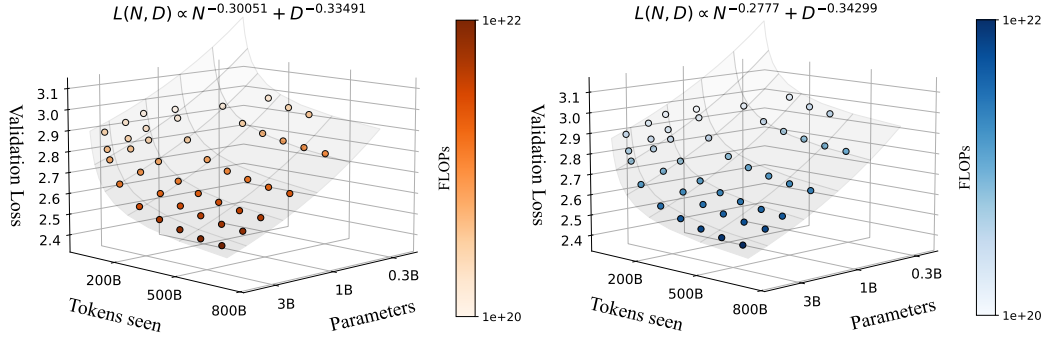


Figure 2: Scaling laws for early-fusion and late-fusion native multimodal models. Each point represents a model (300M to 3B parameters) trained on varying number of tokens (250M to 400B). We report the average cross-entropy loss on the validation sets of interleaved (Obelics), Image-caption (HQITP),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

and number of training tokens. Consistent with prior studies on LLM scaling [Hoffmann et al., 2022], 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.

Data type	dataset	#samples	sampling prob.
Image-Caption	DFN [Fang et al., 2023]	2B	27%
	COYO [Byeon et al., 2022]	600M	11.25%
	HQITP	400M	6.75%
Interleaved	Obelics [Laurençon et al., 2024a]	141M Docs	45%
Text	DCLM [Li et al., 2024b]	6.6T Toks	10%

Table 2: Pre-training data mixture. Unless otherwise specified, the training mixture contains 45%, 45% and 10% of image captions, interleaved documents and text-only data.

2.3 Experimental setup

Our models are based on the autoregressive transformer architecture [Vaswani, 2017] with SwiGLU FFNs [Shazeer, 2020] and QK-Norm [Dehghani et al., 2023] following Li et al. [2024b]. In early-fusion models, image patches are linearly projected to match the text token dimension, while late-fusion follows the CLIP architecture [Radford et al., 2021]. 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 [Li et al., 2024b], Obelics [Laurençon et al., 2024a], DFN [Fang et al., 2023], COYO [Byeon et al., 2022], and a private collection of High-Quality Image-Text Pairs (HQITP) (see table 2). 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) [Zhao et al., 2023], activation checkpointing, and gradient accumulation. We also use sequence packing for the image captioning dataset to reduce the amount of padded tokens. Similar to previous

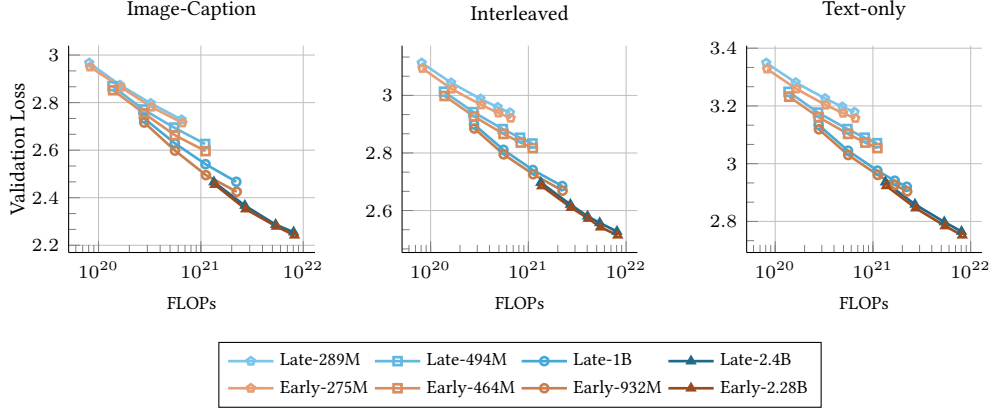


Figure 3: Early vs late fusion: scaling training FLOPs. We compare early and late fusion models when scaling both the number of model parameters and the number of training tokens. Overall, early fusion shows a slight advantage, especially at smaller model sizes, and the gap decreases when scaling the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

works [Hoffmann et al., 2022; Aghajanyan et al., 2023; Abnar et al., 2025], we evaluate performance on a 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 § 3.3, 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 § 3.4.

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. [2024], 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 Equation (1), we apply the L-BFGS algorithm [Nocedal, 1980] and Huber loss [Huber, 1992] (with $\delta = 10^{-3}$), performing a grid search over initialization ranges.

$L \propto E + \frac{1}{N^\alpha} + \frac{1}{D^\beta}$	$N \propto C^a$	$D \propto C^b$			$L \propto C^c$		$D \propto N^d$	
Model	Data	E	α	β	a	b	c	d
GPT3 [Brown et al., 2020]	Text	–	–	–	–	–	-0.048	
Chinchilla [Hoffmann et al., 2022]	Text	1.693	0.339	0.285	0.46	0.54	–	
NMM (early-fusion)	Text	2.222	0.308	0.338	0.525	0.477	-0.042	0.909
	Image-Caption	1.569	0.311	0.339	0.520	0.479	-0.061	0.919
	Interleaved	1.966	0.297	0.338	0.532	0.468	-0.046	0.879
	AVG	1.904	0.301	0.335	0.526	0.473	-0.049	0.899
NMM (late-fusion)	AVG	1.891	0.290	0.338	0.636	0.462	-0.049	0.673
Sparse NMM (early-fusion)	AVG	2.158	0.710	0.372	0.361	0.656	-0.047	1.797

Table 3: 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).

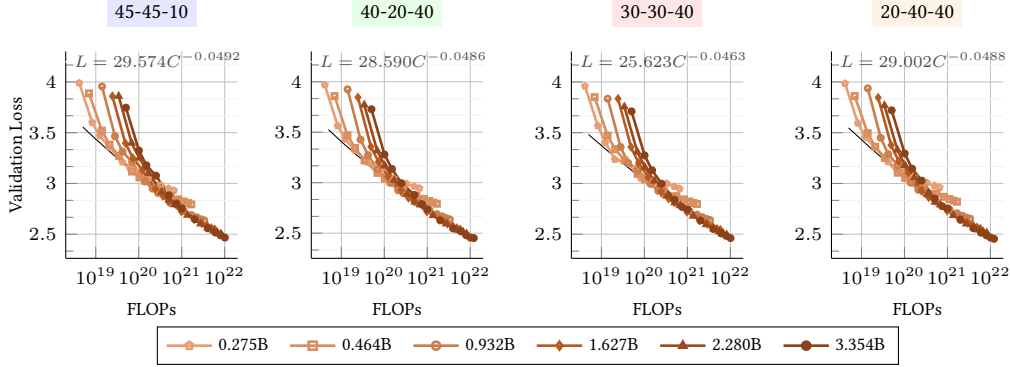


Figure 4: Scaling laws with different training mixtures. Early-fusion models follow similar scaling trends when changing the pretraining mixtures. However, increasing the image captions leads to a higher scaling exponent norm (see table 4).

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 3). For instance, the model achieves a higher rate of improvement for image-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 fig. 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 fig. 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 configurations, 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 [Brown et al., 2020] 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. [2022] ($\alpha =$

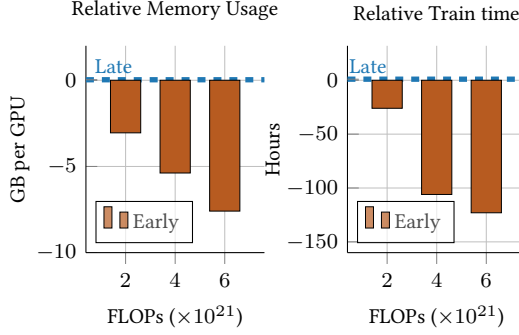


Figure 5: Early vs late: pretraining efficiency. Early-fusion is faster to train and consumes less memory. Models are trained on 16 H100 GPUs for 160k steps (300B tokens).

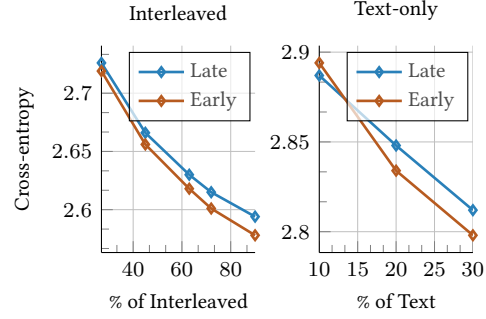


Figure 6: Early vs late fusion: changing the training mixture. We vary the training mixtures and plot the final training loss. Early fusion models attain a favorable performance when increasing the proportion of interleaved documents and text-only data.

	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 4: Scaling laws for different training mixtures. Early-fusion models. C-I-T refer to image-caption, interleaved and text

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 Hoffmann et al. [2022], 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, 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 fig. 1 (right), 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 fig. 5, 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.

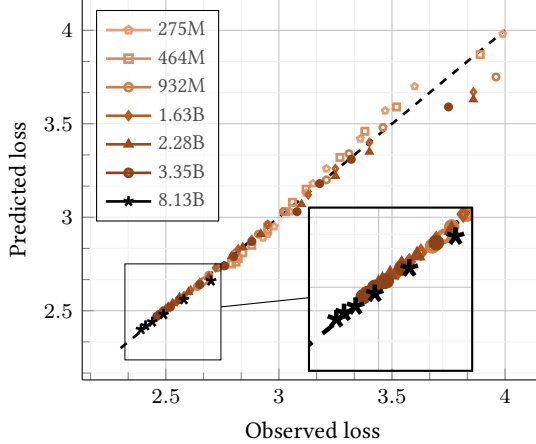


Figure 7: Observed vs predicted loss. We visualize the loss predicted by our scaling laws eq. (1) and the actual loss achieved by each run. We can reliably predict the performance of models larger (8B params) than those used to fit the scaling laws.

3.2 Scaling laws evaluation

For each model size and number of training tokens, we compute the loss using the estimated functional form in eq. (1) and compare it to the actual loss observed in our runs. Figure 7 and Table 5 visualizes these comparisons, showing that our estimation is highly accurate, particularly for lower loss values and larger FLOPs. We also assess our scaling laws in an extrapolation setting, predicting performance beyond the model sizes used for fitting. Notably, our approach estimates the performance of an 8B model with reasonable accuracy.

Additionally, we conduct a sensitivity analysis using bootstrapping. Specifically, we sample P points with replacement (P being the total number of trained models) and re-estimate the scaling law coefficients. This process is repeated 100 times, and we report the mean and standard deviation of each coefficient. Table 6 shows that our estimation is more precise for β than for α , primarily due to the smaller number of model sizes relative to the number of different token counts used to derive the scaling laws.

3.3 Scaling laws for different data mixtures

We investigate how variations in the training mixture affect the scaling laws of native multi-modal models. To this end, we study four different mixtures that reflect common community practices [Laurençon et al., 2024a; McKinzie et al., 2025; Zhang et al., 2024; Lin et al., 2024b], 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, fig. 4 shows that different mixtures follow similar scaling trends; however, the scaling coefficients vary depending on the mixture (table 4). 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 fig. 6.

Parameter	MSE	R2	MAE (%)
Held-in	0.0029	0.9807	0.8608
Held-out	0.0004	0.9682	0.5530

Table 5: Scaling laws prediction errors. We report the mean square error, R2 and mean absolute error for the loss prediction for held-in and held-out (8B model) data.

Parameter	Avg	Std
E	1.80922	0.33811
α	0.29842	0.10101
β	0.33209	0.02892
a	0.54302	0.08813
b	0.48301	0.05787
d	0.92375	0.23296

Table 6: Scaling laws sensitivity. We report the mean and standard deviation after bootstrapping with 100 iterations.

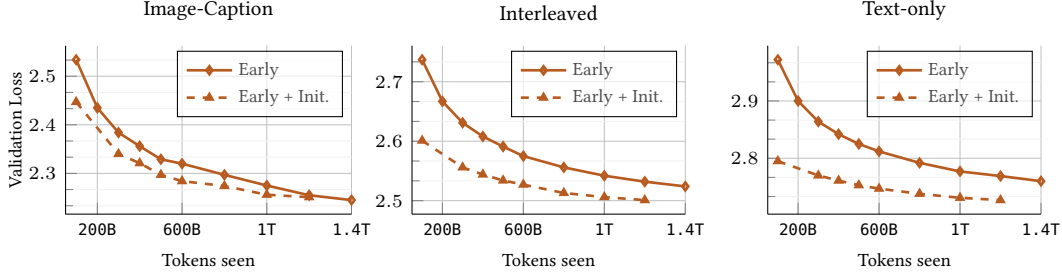


Figure 8: Early native vs initializing from LLMs: initializing from pre-trained models and scaling training tokens. We compare training with and without initializing from DCLM-1B.

3.4 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 [Fang et al., 2023] 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. MoEs 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. [2023], 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 [Shazeer et al., 2017] with a weight of 0.01 to ensure a balanced expert utilization. Following Abnar et al. [2025], 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. fig. 9 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.

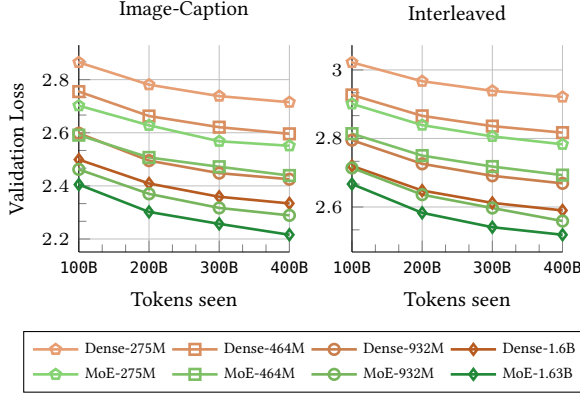


Figure 9: MoE vs Dense: scaling training FLOPs. We compare MoE and dense early-fusion models when scaling both the amount of training tokens and model sizes. MoEs beat dense models when matching the number of active parameters.

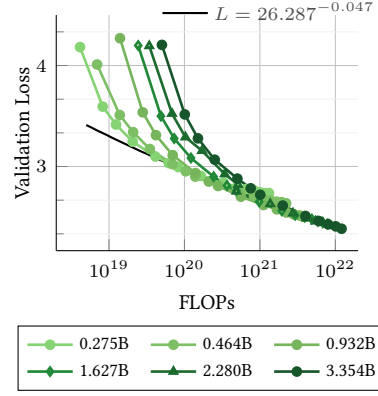


Figure 10: Scaling laws for sparse early-fusion NMMs. We report the final validation loss averaged across interleaved, image-captions and text data.

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 fig. 10. We fit a power law to the convex hull of the lowest loss as a function of compute (FLOPs). Interestingly, the exponent (-0.047) is close to that of dense NMMs (-0.049), indicating that both architectures scale similarly. However, the multiplicative constant is smaller for MoEs (26.287) 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 3 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 [Abnar et al., 2025] 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 their modalities, similar to previous works [Bao et al., 2021; Wang et al., 2022b]. 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 fig. 11.

4.4 Emergence of expert specialization and sharing

We investigate multimodal specialization in MoE architectures. In fig. 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 fig. 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

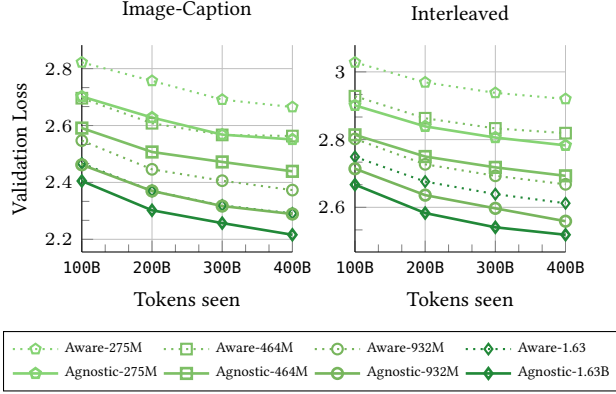


Figure 11: Modality-aware vs modality agnostic routing for sparse NMMs. We compare modality-agnostic routing with modality-aware routing when scaling both the amount of training tokens and model sizes.

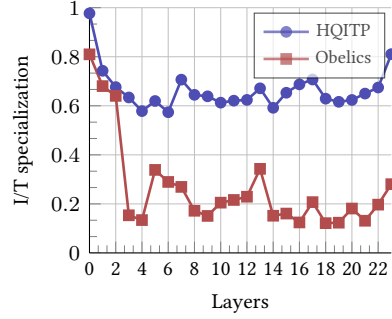


Figure 12: MoE specialization. Entropy-based image/text specialization (see § 4.4) across layers for two data sources: HQITP and Obelics. Both sources exhibit a similar trend: the score decreases in the early layers before increasing again in the final layers.

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 [Shukor and Cord, 2024] 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	VQAv2	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 7: Supervised finetuning on the LLaVA mixture. All models are native at 1.5B scale and pre-trained on 300B tokens.

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 [Liu et al., 2024b] and report accuracy and CIDEr scores across several VQA and captioning tasks. table 7 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 [Team et al., 2023; Hurst et al., 2024]. Crucially, pre-trained vision and language backbones often require surprisingly little adaptation to enable effective cross-modal communication [Tsimpoukelli et al., 2021; Shukor

et al., 2023a; Vallaeyes et al., 2024; Merullo et al., 2023; Koh et al., 2023]. Simply integrating a vision encoder with either an encoder-decoder architecture [Shukor et al., 2023b; Wang et al., 2022a; Lu et al., 2022; Mizrahi et al., 2023] or a decoder-only LLM has yielded highly capable multimodal systems [Laurençon et al., 2024b; Alayrac et al., 2022; Liu et al., 2024b; Wang et al., 2024a; Xue et al., 2024; Chen et al., 2024b; Zhu et al., 2024; Abdin et al., 2024; Dai et al., 2024; Beyer et al., 2024; Moon et al., 2024]. This late-fusion approach, where modalities are processed separately before being combined, is now well-understood, with established best practices for training effective models [Laurençon et al., 2024a; McKinzie et al., 2025; Zhang et al., 2024; Lin et al., 2024b]. In contrast, early-fusion models [Bavishi et al., 2023; Team, 2024; Diao et al., 2024], which combine modalities at an earlier stage, remain relatively unexplored, with only a limited number of publicly released models [Bavishi et al., 2023; Diao et al., 2024]. Unlike [Diao et al., 2024; Team, 2024], 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 [Team et al., 2023] rather than adapting LLMs to accommodate additional modalities. Due to the high cost of training such models, they remain relatively under-explored, with most relying on late-fusion architectures [Huang et al., 2023; Yu et al., 2022]. Some multimodal models trained from scratch [Aghajanyan et al., 2022; Team, 2024; Wang et al., 2024c] relax this constraint by utilizing pre-trained image tokenizers such as [Esser et al., 2021; van den Oord et al., 2017] 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 [Kaplan et al., 2020; Hoffmann et al., 2022] 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 [Krajewski et al., 2024; Clark et al., 2022; Wang et al., 2024b]. Scaling laws have also been observed across various domains, including image models [Fini et al., 2024], video models [Rajasegaran et al., 2025], protein LLMs [Cheng et al., 2024], and imitation learning [Pearce et al., 2024]. However, few studies have investigated scaling laws for multimodal models. Notably, Aghajanyan et al. [2023] 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). Mixture of Experts [Shazeer et al., 2017] enables scaling model capacity by decoupling model size from per-sample compute. This is done through sparsely activating a small number of parameters. This approach has led to large sparse models that rival dense counterparts while being more efficient during training and inference [Fedus et al., 2022; Sun et al., 2024; Jiang et al., 2024; Liu et al., 2024a; Wei et al., 2024]. Many studies have explored improving MoE LLMs across various aspects, such as load balancing, routing, stability, scaling, and granularity [Lewis et al., 2021; Zoph et al., 2022; Lepikhin et al., 2020]. However, there is limited research on adopting MoEs for multimodal models, with some work focusing on contrastive image-text models [Mustafa et al., 2022] and late-fusion multimodal LLMs [Lin et al., 2024a; Li et al., 2024a]. Additionally, some studies investigate predefined expert routing, where certain parameters are reserved to process specific modalities Bao et al. [2021]; Chen et al. [2024a]; Shen et al. [2023]. We focus on studying MoEs for native early-fusion models rather than proposing new architectures.

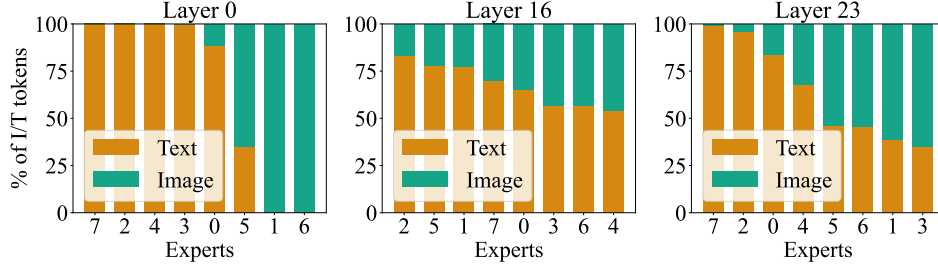


Figure 13: MoE specialization frequency. Percentage of text and image tokens routed to each expert on interleaved data from Obelics. Experts are ordered for better visualization. The first layer shows the highest amount of unimodal experts.

7 Discussion and Limitations

Scaling laws for multimodal data mixtures. Our scaling laws study spans different model configurations and training mixtures. While results suggest that the scaling law coefficients remain largely consistent across mixtures, a broader exploration of mixture variations is needed to validate this observation and establish a unified scaling law that accounts for this factor.

Scaling laws and performance on downstream tasks. Similar to previous scaling law studies, our analysis focuses on pretraining performance as measured by the validation loss. However, the extent to which these findings translate to downstream performance remains an open question and requires further investigation.

Extrapolation to larger scales. The accuracy of scaling law predictions improves with increasing FLOPs appendix C. Furthermore, we validate our laws when extrapolating to larger model sizes (§ 3.2). However, whether these laws can be reliably extrapolated to extremely large model sizes remains an open question.

High resolution and early-fusion models. Training early-fusion models with high-resolution inputs leads to a significant increase in vision tokens. While pooling techniques have been widely adopted for late-fusion models, alternative approaches may be necessary for early fusion. Given the similarity of early-fusion models to LLMs, it appears that techniques for extending context length could be beneficial.

Scaling laws for multimodal MoEs models. For MoEs, we consider only a single configuration (top-1 routing with 8 experts). We found this configuration to work reasonably well in our setup, and follow a standard MoEs implementation. However, the findings may vary when optimizing more the MoE architecture or exploring different load-balancing, routing strategies or different experts implementations.

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

- Marah Abdin, Jyoti Aneja, Hany Awadalla, 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.
- 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.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- 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.
- 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.
- 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.
- Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmoe: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.
- Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşlılar. Introducing our multimodal models, 2023.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 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.
- 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, 2024a.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

- 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, 2024b.
- Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. *bioRxiv*, 2024.
- 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.
- Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoneybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- 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.
- 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.
- 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.
- Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- 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.
- Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.
- 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.
- Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrissi 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.
- 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.
- 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.
- 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.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

- Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.
- 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.
- Peter J. Huber. *Robust Estimation of a Location Parameter*, pages 492–518. Springer New York, New York, NY, 1992.
- 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.
- 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.
- 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.
- 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.
- Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, pages 25125–25148. PMLR, 2024.
- 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.
- 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, 2024a.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024b.
- 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.
- 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.
- 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*, 2024a.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024b.
- 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*, 2024a.
- 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, 2024b.

- 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*, 2024a.
- 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, 2024b.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- 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.
- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi 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.
- 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.
- 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.
- Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Anymal: 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.
- 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.
- Jorge Nocedal. Updating quasi newton matrices with limited storage. *Mathematics of Computation*, 35(151): 951–958, 1980.
- 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.
- 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.
- 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.
- Jathushan Rajasegaran, Ilija Radosavovic, Rahul Ravishankar, Yossi Gandelsman, Christoph Feichtenhofer, and Jitendra Malik. An empirical study of autoregressive pre-training from videos. *arXiv preprint arXiv:2501.05453*, 2025.
- Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, 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.
- 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.
- Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen llms to multimodal inputs. *arXiv preprint arXiv:2405.16700*, 2024.

- 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, 2023a.
- 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*, 2023b.
- 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.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- 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.
- 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.
- Th  ophane Valla  ys, Mustafa Shukor, Matthieu Cord, and Jakob Verbeek. Improved baselines for data-efficient perceptual augmentation of llms. *arXiv preprint arXiv:2403.13499*, 2024.
- Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.
- A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- 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, 2022a.
- 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*, 2024a.
- 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, 2024b. Association for Computational Linguistics.
- 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*, 2022b.
- 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*, 2024c.
- Tianwen Wei, Bo Zhu, Liang Zhao, Cheng Cheng, Biye Li, Weiwei L  , 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.
- 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.
- 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.

- 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.
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023.
- Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024.
- 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.
- 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.
- Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

Appendices

A	Experimental setup	21
B	Late vs early fusion	22
B.1	Scaling FLOPs	22
B.2	Changing the training data mixture	22
B.3	Scaling image resolution is in favor of early-fusion	23
B.4	Early-fusion is consistently better when matching the late-fusion model size . .	24
B.5	Different late-fusion configuration	24
B.6	Initializing from LLM and CLIP	25
C	Scaling laws	26
C.1	Fitting $L = F(N, D)$	26
C.2	Fitting $N \propto C^a$, $D \propto C^b$ and $D \propto N^d$	26
C.3	Fitting $L \propto C^c$	26
C.4	Scaling laws for different target data type	27
C.5	Scaling laws for different training mixtures	28
C.6	Scaling laws evaluation and sensitivity	28
C.7	Scaling laws for sparse NMMs.	28
D	Mixture of experts and modality-specific specialization	29

A Experimental setup

In Table 8, we show the pre-training hyperparameters for different model configurations used to derive the scaling laws. The number of parameters ranges from 275M to 3.7B, with model width increasing accordingly, while the depth remains fixed at 24 layers. Learning rates vary by model size, decreasing as the model scales up. Based on empirical experiments and estimates similar to [McKinzie et al., 2025], we found these values to be effective in our setup. Training is optimized using a fully decoupled AdamW optimizer with momentum values $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of $1e-4$. The batch size is set to 2k samples, which account for 2M tokens, given a 1k context length. Gradient clipping is set to 1.0, with a maximum warmup duration of 5k iterations, adjusted for shorter training runs: 1k and 2.5k warmup steps for models trained between 1k–4k and 5k–15k steps, respectively. For MoEs, we found that a longer warmup is significantly better, so we adopt a 2.5k warmup for all runs under 20k steps. We use a constant learning rate schedule with cooldown during the final 20% of training, gradually reducing to zero following an inverse square root schedule. For vision processing, image inputs are divided into (14, 14) patches, with augmentations including Random Resized Crop (resizing images to 224px with a scale range of [0.4, 1.0]) and Random Horizontal Flip with a probability of 0.5. We train our models on mixture of interleaved, image captions and text only data Table 2. For late fusion models, we found that using smaller learning rate for the vision encoder significantly boost the performance Table 10, and when both the encoder and decoder are initialized (Appendix B.6) we found that freezing the vision encoder works best Table 9.

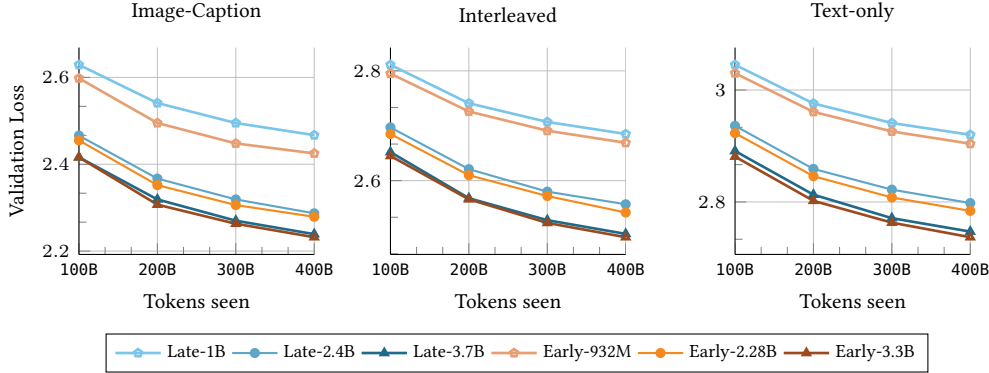


Figure 14: Early vs late fusion: scaling training FLOPs. We compare early and late fusion models when scaling both the model size and the number of training tokens. The gap decreases mainly due to scaling models size.

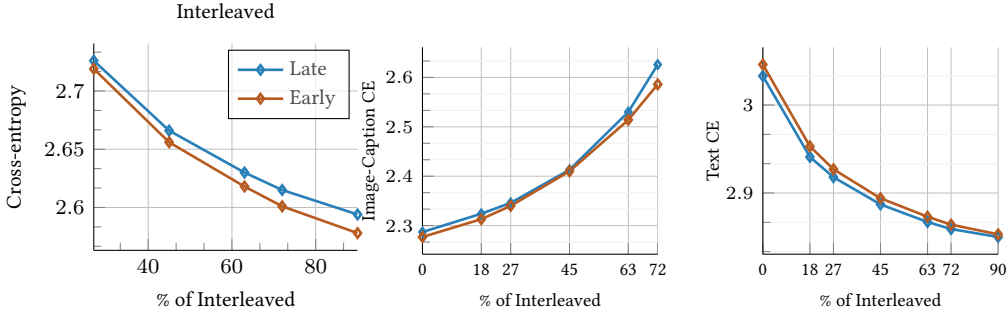


Figure 15: Early vs late fusion: changing the training mixture. We vary the training mixtures and plot the final training loss. Early fusion models become better when increasing the proportion of interleaved documents. Early and late fusion has 1.63B and 1.75B parameters respectively.

Early-fusion						
Params	275M	468M	932M	1.63B	2.28B	3.35B
width	800	1088	1632	2208	2624	3232
depth				24		
Learning rate	1.5e-3	1.5e-3	5e-4	4.2e-4	4e-4	3.5e-4
Late-fusion						
Params	289M	494M	1B	1.75B	2.43B	3.7B
vision encoder width	384	512	768	1024	1184	1536
vision encoder depth				24		
width	768	1024	1536	2048	2464	3072
depth				24		
Learning rate	1.5e-3	1.5e-3	5e-4	4.2e-4	3.8e-4	3.3e-4
Early-fusion MoEs						
Active Params	275M	468M	932M	1.63B	2.28B	3.35B
width	800	1088	1632	2208	2624	3232
depth				24		
Learning rate	1.5e-3	1.5e-3	5e-4	4.2e-4	4e-4	3.5e-4
Training tokens	2.5B-600B					
Optimizer	Fully decoupled AdamW [Loshchilov and Hutter, 2017]					
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$					
Minimum Learning rate	0					
Weight decay	1e-4					
Batch size	2k					
Patch size	(14, 14)					
Gradient clipping	1.0					
MAximum Warmup iterations	5k					
Augmentations:						
RandomResizedCrop						
size	224px					
scale	[0.4, 1.0]					
RandomHorizontalFlip	$p = 0.5$					

Table 8: Pre-training hyperparameters We detail the hyperparameters used for pre-training different model configurations to derive scaling laws.

B Late vs early fusion

This section provides additional comparison between early and late fusion models.

B.1 Scaling FLOPs

Figure 14 compares early-fusion and late-fusion models when scaling FLOPs. Specifically, for each model size, we train multiple models using different amounts of training tokens. The performance gap between the two approaches mainly decreases due to increasing model sizes rather than increasing the number of training tokens. Despite the decreasing gap, across all the models that we train, early-fusion consistently outperform late-fusion.

B.2 Changing the training data mixture

We analyze how the performance gap between early and late fusion models changes with variations in the training data mixture. As shown in Figure 16 and Figure 15, when fixing the model size, increasing the ratio of text and interleaved data favors early fusion. Interestingly, the gap remains largely unchanged for other data types. We also observe interference effects between different data

Vision encoder lr scaler	Interleaved (CE)	Image-Caption (CE)	Text (CE)	AVG (CE)	AVG (SFT) (Acc)
1	2.521	2.15	2.867	2.513	43.49
0.1	2.502	2.066	2.862	2.477	52.27
0.01	2.502	2.066	2.859	2.476	53.76
0.001	2.513	2.066	2.857	2.479	–
0 (frozen)	2.504	2.061	2.856	2.474	54.14

Table 9: Vision encoder scaler. Freezing the vision encoder works best when initializing late-fusion models with pre-trained models.

Vision encoder lr scaler	Interleaved (CE)	Image-Caption (CE)	Text (CE)	AVG (CE)	AVG (SFT) (Acc)
0.1	2.674	2.219	3.072	2.655	34.84
0.01	2.672	2.197	3.071	2.647	38.77
0.001	2.674	2.218	3.073	2.655	38.46

Table 10: Vision encoder scaler. Reducing the learning rate for the vision encoder is better when training late-fusion models from scratch.

types. Specifically, increasing the amount of interleaved data negatively impacts performance on image captions and vice versa. Additionally, increasing the proportion of text-only data slightly improves interleaved performance but increases loss on image captions. Overall, we find that text-only and interleaved data are correlated across different setups.

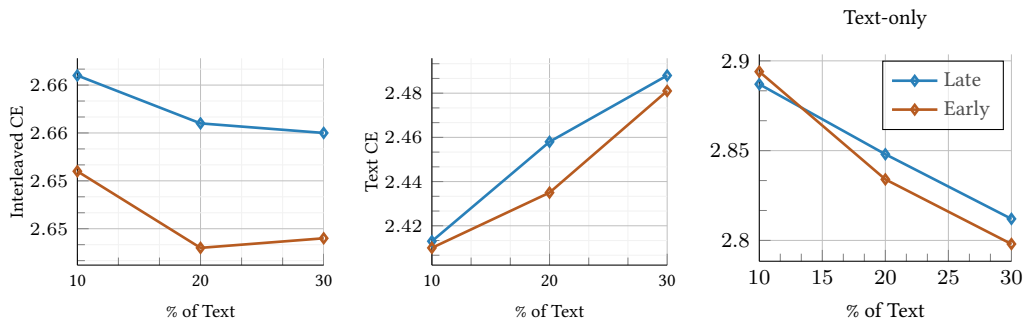


Figure 16: Early vs late fusion: changing the amount of text-only data in the training mixture (isoFLOPs). We vary the ratio of text-only data and plot the final training loss. The gap increases with the text data ratio in favor of early fusion model. Early fusion has 1.63B parameters and late fusion 1.75B parameters.

B.3 Scaling image resolution is in favor of early-fusion

We examine how both architectures perform with varying image resolution. We fix the number of model parameters to 1.63B and 1.75B for early and late fusion respectively. All models are trained for 100K steps or 200B tokens. Since the patch size remains constant, increasing the resolution results in a higher number of visual tokens. For all resolutions, we maintain the same number of text tokens. As shown in Figure 17, the early-fusion model consistently outperforms the late-fusion model across resolutions, particularly for multimodal data, with the performance gap widening at higher resolutions. Additionally, we observe that the loss on text and interleaved data increases as resolution increases.

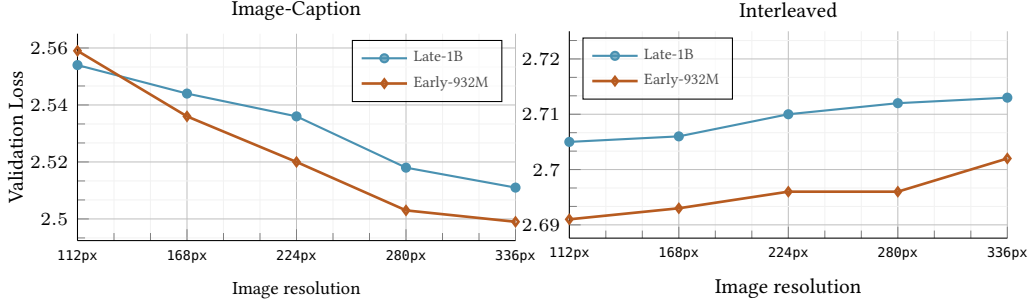


Figure 17: Early vs late fusion: training with different image resolutions (isoFLOPs). For the same training FLOPs we vary the image resolution (and thus the number of image tokens) during training and report the final training loss. Increasing resolution, hurts the performance on text and interleaved documents, while helping image captioning. The gap stays almost the same on text and interleaved data while slightly increase on image captioning in favor of early fusion.

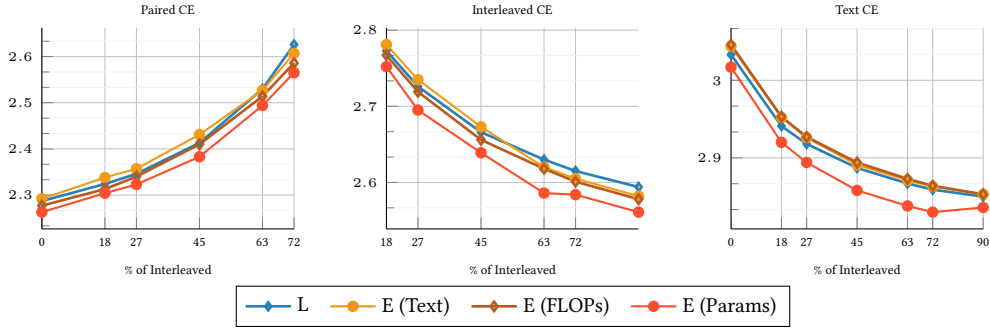


Figure 18: Early vs late fusion: changing the training mixture and early-fusion configuration. We vary the training mixtures and plot the final training loss for different configuration of early fusion models. For the same number of total parameters early fusion consistently outperform late fusion.

B.4 Early-fusion is consistently better when matching the late-fusion model size

In this section, we compare the late-fusion model with different configurations of early-fusion one. Specifically, we train early-fusion models that match the late-fusion model in total parameters (Params), text model size (Text), and FLOPs (FLOPs), assuming 45-45-10 training mixture. As shown in Figure 18, early fusion consistently outperforms late fusion when normalized by total parameters, followed by normalization by FLOPs. When matching the text model size, early fusion performs better at higher ratios of interleaved data.

B.5 Different late-fusion configuration

We examine how this scaling changes with different late-fusion configurations. Instead of scaling both the vision and text models equally, as done in the main paper, we fix the vision encoder size to 300M and scale only the text model. Figure 19 shows that late-fusion models lag behind at smaller model sizes, with the gap closing significantly as the text model scales. This suggests that allocating more parameters to shared components is more beneficial, further supporting the choice of early-fusion models.

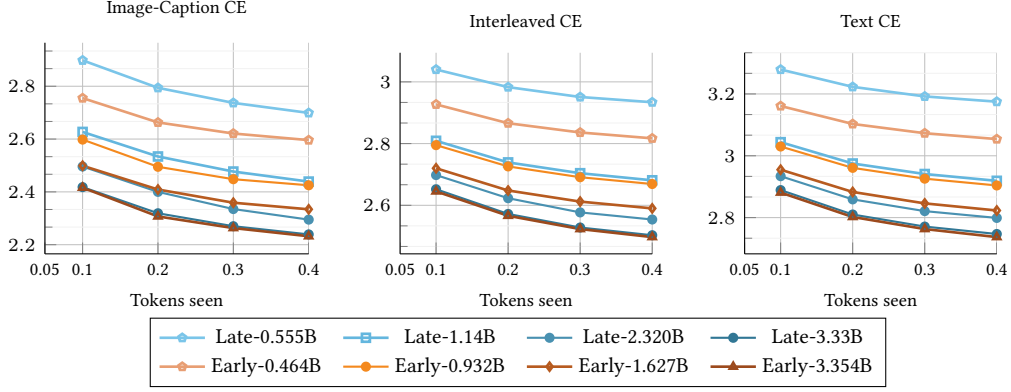


Figure 19: Early vs late fusion: scaling training FLOPs while fixing the vision encoder size. We compare early and late fusion models when scaling both the amount of training tokens and model sizes. For late fusion models, we fix the vision encoder size (300M) and scale the text model (250M, 834M, 2B, 3B). The gap between early and late fusion gets tighter when scaling the text model.

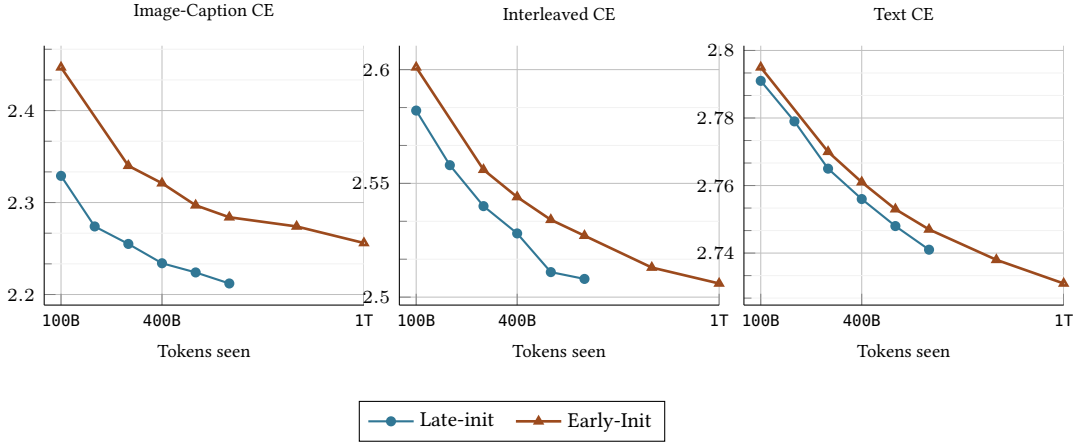


Figure 20: Early vs late fusion when initializing the encoder and decoder. Early-fusion can match the performance of late-fusion models when trained for longer. However, the gap is bigger on image-caption data.

B.6 Initializing from LLM and CLIP

We study the case where both late and early fusion models are initialized from pre-trained models, specifically DCLM-1B [Li et al., 2024b] and CLIP-ViT-L [Radford et al., 2021] for late fusion. Interestingly, Figure 20 shows that for text and interleaved multimodal documents, early fusion can match the performance of late fusion when trained for longer. However, closing the gap on image caption data remains more challenging. Notably, when considering the overall training cost, including that of pre-trained models, early fusion requires significantly longer training to compensate for the vision encoder’s pretraining cost.

C Scaling laws

C.1 Fitting $L = F(N, D)$

Following [Hoffmann et al., 2022], we determine the parameters that minimize the following objective across all our runs i :

$$\min_{a,b,e,\alpha,\beta} \sum_i \text{Huber}_\delta (\text{LSE}(a - \alpha \log N_i, b - \beta \log D_i, e) - \log L_i), \quad (2)$$

We perform this optimization across various initialization ranges and select the parameters that achieve the lowest loss across all initializations. Specifically, our grid search spans $\{0, 0.5, 2.5\}$ for α and β , $\{0, 5, 10, \dots, 30\}$ for a and b , and $\{-1, -0.5, 1, 0.5\}$ for e . We use the L-BFGS algorithm with $\delta = 1e - 3$.

C.2 Fitting $N \propto C^a$, $D \propto C^b$ and $D \propto N^d$

While these equations have a closed-form solution [Hoffmann et al., 2022] for early-fusion models that can be derived from Equation (1), this is not the case for late-fusion models without specifying either the vision encoder or text model size. To ensure a fair comparison, we derive these equations for both models, by performing linear regression in log space. We found that the regression is very close to the coefficient found with closed-form derivation Table 11. For instance, to derive $N = K_a C^a$, given a FLOP budget C and a set of linearly spaced tokens D_i ranging from 10B to 600B, we compute the model size for each D_i as $N_i = \frac{C}{6D}$ for early fusion and $N_i = \frac{C}{6D} + 0.483 * N_v$ for late fusion (for the 45-45-10 mixture, $D_v = 0.544D$, thus $C = 6D(0.544N_v + N_i)$). We then apply Equation (1) to obtain the loss for each model size and select N that has the minimum loss. We repeat this for all FLOP values corresponding to our runs, resulting in a set of points (C, N_{opt}) that we use to regress a and K_a . We follow a similar procedure to find b and d . For late-fusion models, we regress a linear model to determine N_v given N . Notably, even though we maintain a fixed width ratio for late-fusion models, this approach is more accurate, as embedding layers prevent a strictly fixed ratio between text and vision model sizes. We present the regression results in Figure 21.

Model	a	b	d	n	dn
Closed form	0.52649	0.47351	0.89938	1.11188	-0.05298
Regression	0.52391	0.47534	0.90052	1.10224	-0.04933

Table 11: Scaling laws parameters for early-fusion. Doing regression to derive the scaling laws coefficients leads to very close results to using the closed-form solution.

C.3 Fitting $L \propto C^c$

To determine the relationship between the final model loss and the compute budget C , we begin by interpolating the points corresponding to the same model size and compute the convex hull that covers the minimum loss achieved by all runs for each FLOP. This results in a continuous mapping from the FLOPs to the lowest loss. We consider a range of FLOPs, excluding very small values ($\leq 3e^{19}$), and construct a dataset of (C, L) for linearly spaced compute C . Using this data, we find the linear relationship between L and C in the log space and deduce the exponent c . We visualize the results in Figure 24.

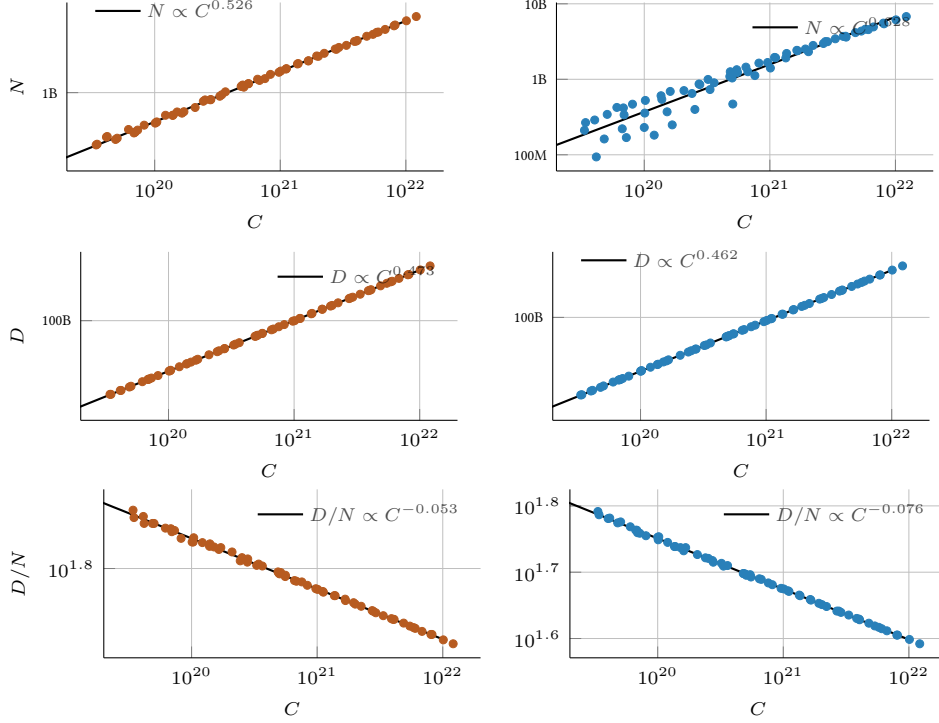


Figure 21: Regression results of the scaling laws coefficients. our estimation of the scaling coefficients is close to the closed form solution.

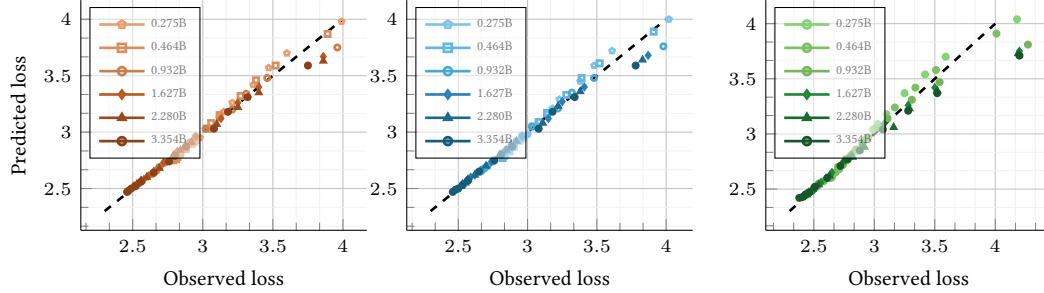


Figure 22: Observed vs predicted loss. We visualize the loss predicted by our scaling laws (Equation (1)) and the actual loss achieved by each run.

C.4 Scaling laws for different target data type

In Figure 25, we derive the scaling laws for different target data types. In general, we observe that the model learns image captioning faster than interleaved data, as indicated by the higher absolute value of the scaling exponent (e.g., 0.062 vs 0.046), despite using the same data ratio for captioning and interleaved data (45% each). Additionally, we find that the model learns more slowly on text-only data, likely due to the smaller amount of text-only data (10%). Across model configurations, we find that early fusion scales similarly to late fusion on image captioning but has a lower multiplicative constant (49.99 vs 47.97). For MoEs, the model learns faster but exhibits a higher multiplicative constant. On text and interleaved data, early and late fusion models scale similarly and achieve comparable performance. However, MoEs demonstrate better overall performance while learning slightly more slowly.

C.5 Scaling laws for different training mixtures

We investigate how the scaling laws change when modifying the training mixtures. Specifically, we vary the ratio of image caption, interleaved, and text-only data and report the results in Figure 26. Overall, we observe similar scaling trends, with only minor changes in the scaling coefficients. Upon closer analysis, we find that increasing the ratio of a particular data type in the training mixture, leads to a corresponding increase in its scaling exponent. For instance, increasing the ratio of image captions from 30% to 40% raises the absolute value of the exponent from 0.056 to 0.061. However, for text-only data, we do not observe significant changes in the scaling coefficients when varying its proportion in the training mixture.

C.6 Scaling laws evaluation and sensitivity

For each model size and number of training tokens, we compute the loss based on the estimated functional form in Equation (1) and compare it with the actual loss achieved by our runs. We visualize these points in Figure 22, demonstrating that our estimation is highly accurate, particularly for lower loss values, and hence for larger FLOPs. Additionally, we perform a sensitivity analysis using bootstrapping. Specifically, we sample with replacement P points (P being equal to the total number of trained models) and re-estimate the scaling law coefficients. This process is repeated 100 times, and we report the average and standard deviation of each coefficient. Table 12 shows that our estimation is more precise for β compared to α , primarily due to the smaller number of model sizes relative to the number of different token counts used to derive the scaling laws.

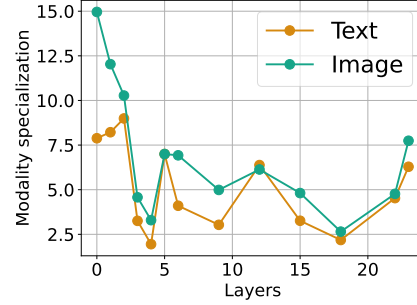


Figure 23: Modality-specific specialization. We visualize the experts specialization to text and image modalities. Models are evaluated on Obelics.

Model	E	α	β	a	b	d
Avg	1.80922	0.29842	0.33209	0.54302	0.48301	0.92375
Std	0.33811	0.10101	0.02892	0.08813	0.05787	0.23296

Table 12: Scaling laws sensitivity. We report the mean and standard deviation after bootstrapping with 100 iterations.

C.7 Scaling laws for sparse NMMs.

Similar to dense models, we fit a parametric loss function (Equation (1)) to predict the loss of sparse NMMs based on the number of parameters and training tokens, replacing the total parameter count with the number of active parameters. While incorporating sparsity is standard when deriving scaling laws for MoEs [Wang et al., 2024b; Krajewski et al., 2024; Abnar et al., 2025], we focus on deriving scaling laws specific to the sparsity level used in our MoE setup. This yields coefficients that are implicitly conditioned on the sparsity configuration.

We also experiment with a sparsity-aware formulation of the scaling law as proposed in [Abnar et al., 2025], and observe consistent trends (Table 13). In particular, the exponents associated with model size (N) are substantially larger than those for training tokens (β), reinforcing the importance of scaling model size in sparse architectures. Additionally, we observe that the terms governing the scaling of active parameters decompose into two components.

$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$			vs		$L(N, D, S) = \frac{A}{N^\alpha} + \frac{B}{D^\beta} + \frac{C}{(1-S)^\lambda} + \frac{d}{(1-S)^\delta N^\gamma} + E$						
Model	E	A	B	α	β	λ	δ	γ	C	d	
$L(N, D)$ Equation (1)	2.158	381773	4659	0.710	0.372	–	–	–	–	–	
$L(N, D, S)$ [Abnar et al., 2025]	1.0788	1	4660	0.5890	0.3720	0.2	0.2	0.70956	1.0788	381475	

Table 13: Scaling laws for sparse native multimodal models. Higher exponent for active parameters.

D Mixture of experts and modality-specific specialization

We investigate multimodal specialization in MoE architectures. We compute a specialization score as the average difference between the number of text/images tokens assigned to each expert and a uniform assignment ($1/E$). Additionally, we visualize the normalized number of text and image tokens assigned to each expert across layers. Figure 23 shows clear modality-specific experts, particularly in the early layers. Furthermore, the specialization score decreases as the number of layers increases but rises again in the very last layers. This suggests that early and final layers require more modality specialization compared to mid-layers. Additionally, we observe several experts shared between text and image modalities, a phenomenon not present in hard-routed or predefined modality-specific experts.

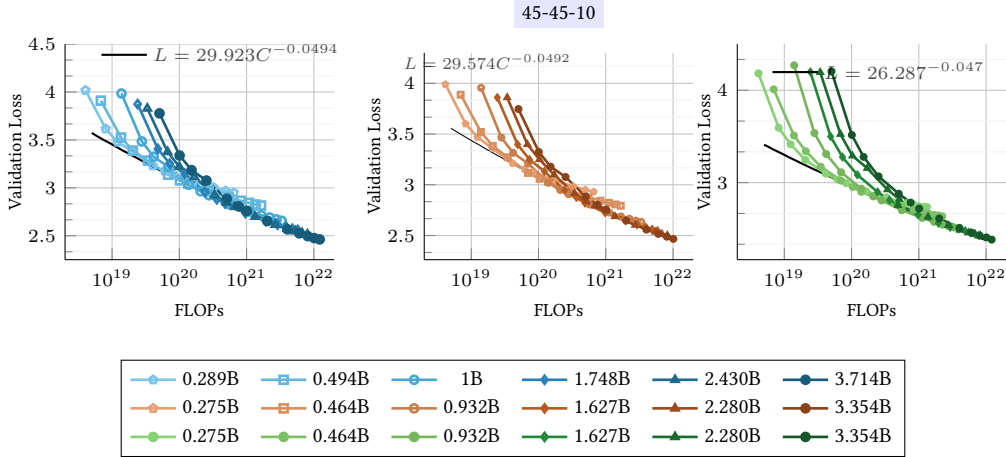


Figure 24: Scaling laws for native multimodal models. From left to right: late-fusion (dense), early-fusion (dense) and early-fusion MoEs. The scaling exponents are very close for all models. However, MoEs leads to overall lower loss (smaller multiplicative constant) and takes longer to saturate.

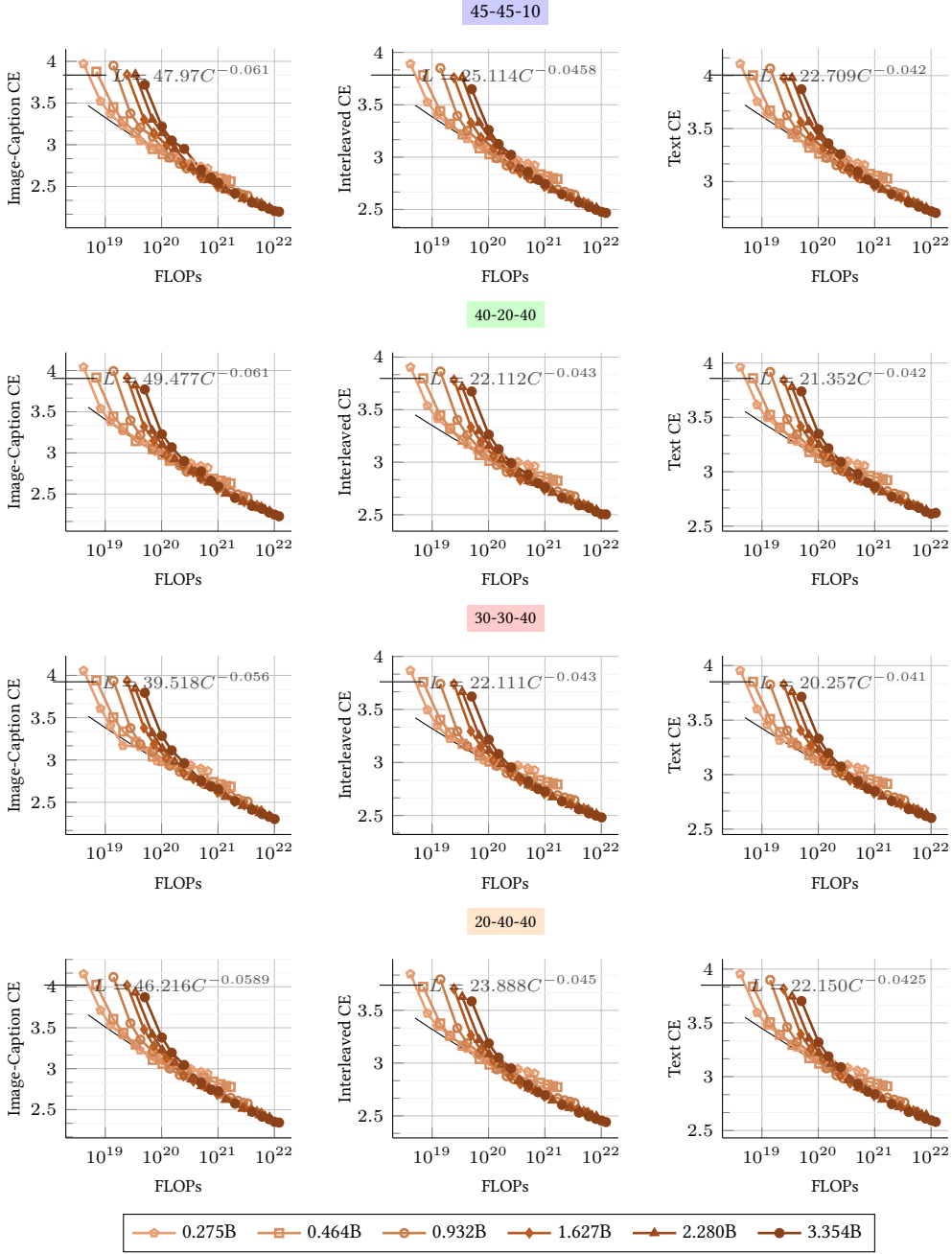


Figure 26: Scaling laws for early-fusion native multimodal models. Our runs across different training mixtures (Image-caption-Interleaved-Text) and FLOPs. We visualize the final validation loss on 3 data types: HQITP (left), Obelics (middle) and DCLM (right).