SLIM AND SPARSE: TOWARDS EFFICIENT UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale multimodal models have achieved remarkable progress in both understanding and generation. Traditionally, these tasks were studied in isolation, resulting in separate architectures. Recent efforts instead pursue unified multimodal models that combine heterogeneous components to support both capabilities within a single framework. However, such models introduce substantial challenges related to architectural redundancy, compute allocation, and efficient scaling. In this work, we conduct a systematic analysis of unified multimodal model components using training-free pruning as a probing methodology, considering both depth pruning and width reduction. Our study reveals that the understanding component, although essential for multimodal reasoning, exhibits notable compressibility in generation tasks. In contrast, the generation components are highly sensitive to compression, with performance degrading sharply even under moderate ratios of depth or width reduction. To address this limitation, we propose a Mixture-of-Experts (MoE) Adaptation, inspired by the dynamic activation patterns observed in hidden neurons. This approach partitions the generation module into multiple experts and enables sparse activation to restore generation quality. We first demonstrate the potential of sparse activation in generation components, and then show that a fully trainable adaptation further enhances performance. As a result, the adapted BAGEL model achieves performance comparable to the full model while activating only about half of the parameters.

1 INTRODUCTION

Large-scale multimodal models have recently achieved remarkable progress in both multimodal understanding (Liu et al., 2023a; Li et al., 2023; Dai et al., 2023; Lu et al., 2024; 2023) and generation (Ramesh et al., 2021; Saharia et al., 2022; Peebles & Xie, 2023). Traditionally, these two tasks were studied in isolation, leading to distinct research trajectories and model families: *understanding-oriented architectures* for vision–language reasoning with textual outputs, and *generative models* designed for image synthesis. While effective for task-specific purposes, this separation stands in contrast to the broader pursuit of Artificial General Intelligence (AGI) (Wei et al., 2022; Bubeck et al., 2023), where a single model is expected to both interpret and generate across modalities in a unified manner.

Motivated by this vision, recent research has shifted toward Unified multimodal models that unify multimodal understanding and generation within a single framework (Deng et al., 2025; Liang et al., 2025; AI et al., 2025). By integrating heterogeneous components such as vision encoders (Dosovitskiy et al., 2021), language backbones (Grattafiori et al., 2024; Yang et al., 2024), and image or audio decoders Peebles & Xie (2023); AI et al. (2025), Unified multimodal models can seamlessly support reasoning tasks and generative tasks in the same system. This paradigm promises more general-purpose multimodal intelligence and has already demonstrated encouraging capabilities across diverse benchmarks.

However, this unification comes at a substantial **cost in efficiency**. Unlike unimodal or task-specific multimodal models (Peebles & Xie, 2023; Liu et al., 2023a), Unified multimodal models must support outputs of different modalities while sharing internal components across tasks. This creates several inefficiencies: 1) **architectural redundancy:** shared modules often house parameters that are only useful for a subset of tasks, leading to under-utilized capacity; 2) **compute allocation challenges**

lenges: the same backbone must simultaneously support reasoning-oriented token processing and high-fidelity generation, yet the compute demands of these tasks differ significantly; and 3) **scaling uncertainty:** as models grow larger, it remains unclear how best to distribute depth and width across understanding versus generation pathways to maximize performance per parameter.

In this work, we conduct a systematic investigation of the components of Unified multimodal models and uncover substantial redundancy from multiple perspectives. To this end, we employ **training-free pruning** as a probing methodology, examining via depth pruning (Gromov et al., 2025) (e.g., dropping transformer blocks or attention layers) and neuron partition (e.g., compressing structured hidden neurons). We begin by analyzing the understanding components, which are the shared modules responsible for processing inputs across different modalities, since they form the backbone of multimodal representation learning and often serve as the foundation for downstream reasoning and generation. Our results show that these understanding components exhibit notable compressibility in multimodal generation tasks, where pruned models can still sustain competitive performance. Furthermore, we observe clear task-specific activation patterns: understanding and generation tasks predominantly activate different model partitions, underscoring the necessity of dynamic pruning for different testing tasks.

However, when compressing the generation components (e.g., image generators), we observe that the quality of generated images drops drastically after either depth pruning or neuron partition. To address this issue, we propose a Mixture-of-Experts (MoE) Adaptation, inspired by the dynamic activation patterns observed across different prompts and diffusion steps. In this approach, neurons in the MLP layers are partitioned into experts, allowing the model to selectively activate subsets of neurons and thereby restore generation quality. We first validate this idea with Expert-Frozen Tuning (EFT), where experts remain frozen while the router and other parameters are optimized to align with sparse activation. This stage already recovers a substantial portion of the lost generation capability. Building on this, we further train the MoE model in a fully end-to-end manner, which delivers additional improvements. As a result, the adapted BAGEL model (Deng et al., 2025) achieves performance comparable to the full model while activating only about half of the neurons.

2 Related Works

Unification of Understanding and Generation Traditionally, multimodal understanding and generation were studied as separate tasks, which in turn gave rise to two distinct streams of multimodal model architectures (Li et al., 2023; Dai et al., 2023). On the one hand, multimodal large language models (MLLMs) extend language models to handle input tokens from multiple modalities. For instance, LLaVA (Liu et al., 2023a) builds upon the LLaMA backbone (Touvron et al., 2023a) by incorporating both text and image tokens, and subsequent multimodal training substantially enhances its ability to perform vision–language understanding tasks such as visual question answering. On the other hand, multimodal generative models typically employ a text encoder to convert natural language into embeddings, which then serve as conditional signals for image generators. Recent advances in diffusion-based architectures, such as DiT (Diffusion Transformers) (Peebles & Xie, 2023), demonstrate that transformer backbones can effectively model the denoising process, while techniques like classifier-free guidance (CFG) (Ho & Salimans, 2022) further improve controllability and fidelity in conditional image synthesis. Despite their separate origins, more recent research has increasingly aimed to unify these two paradigms within a single architecture, enabling models to seamlessly perform both multimodal understanding and generation. For instance, BAGEL (Deng et al., 2025) adopts an interleaved multimodal training paradigm coupled with a mixture-of-transformers design (Liang et al., 2025) that separates understanding and generation modules, while Ming-Omni (AI et al., 2025) employs a Mixture-of-Experts (MoE) backbone with dedicated routing mechanisms and modality-specific decoders to integrate text, vision, audio, and video within a single unified framework.

Model Compression toward Parameter Efficiency Despite the remarkable advances of large language models, the continual growth in their size has introduced substantial redundancy and raised critical challenges for scalability. Network pruning (Cheng et al., 2024) has emerged as an effective technique to identify and alleviate architectural redundancy. For instance, Gromov et al. (2025) demonstrated that many deep layers in large language models are relatively unimportant, and that comparable performance can still be maintained after removing these redundant layers. He et al.

(2024) identified redundancy within attention layers, showing that a large proportion of them can be removed without significantly affecting performance on textual question answering tasks. While the uni-modal compression techniques can be transferred to Vision-Language models that take multi-modal inputs and output the language responses via language models (Sung et al., 2024; He et al., 2025), it is unclear whether such methods still work in Unified models, We take the prior efforts to systematically explore and exploit redundancy in multimodal models, where heterogeneous components play distinct roles. This perspective enables us to design compression strategies better aligned with the unified nature of multimodal understanding and generation.

3 Omni Models Unifying Understanding and Generation

Unified models are large-scale multimodal architectures that aim to unify understanding and generation within a single framework. Unlike traditional multimodal systems, which either focus on reasoning (e.g., vision–language question answering) or on generation (e.g., text-to-image or text-to-speech synthesis), Unified models are designed to support both modalities simultaneously, thereby moving closer to the goal of Artificial General Intelligence (AGI).

Given an Unified model, let $\mathbf{x}$ denote the multimodal input tokens (e.g., text, image, or audio), and $\mathbf{y}$ the target output (e.g., text or image).

Understanding. For *understanding tasks*, the model predicts textual outputs in an auto-regressive manner:

$$p(\mathbf{y}_{\text{und}} \mid \mathbf{x}; \theta_{\text{und}}) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, \mathbf{x}; \theta_{\text{und}}) \,, \tag{1}$$

where θ_{und} denotes the parameters of the understanding component, responsible for both multimodal feature extraction and language modeling.

Generation. For *generation tasks*, the Unified model leverages the understanding component to process an instructional input $\mathbf{x}_{\text{inst}}$ (e.g., text prompt and reference images) , producing conditional features $f_{\text{und}}(\mathbf{x}_{\text{inst}}; \theta_{\text{und}})$. The generative component then synthesizes the output $\mathbf{y}_{\text{gen}}$, typically conditioned on both this representation and an additional generative input (e.g., random noise $\mathbf{z}$ in diffusion models or initial tokens in auto-regressive decoding):

$$\mathbf{y}_{\text{gen}} \sim p(\mathbf{y} \mid f_{\text{und}}(\mathbf{x}_{\text{inst}}; \theta_{\text{und}}), \mathbf{z}; \theta_{\text{gen}}) \,, \tag{2}$$

where θ_{gen} are the parameters of the generative component.

Overall, Unified models unify multimodal understanding and generation through a shared understanding component θ_{und}, whose outputs serve either as predictions for understanding tasks or as instructional signals for non-text generation. For modalities such as images or audio, this shared component is further coupled with modality-specific generators. Given that this unification integrates heterogeneous components with distinct functional roles, we next conduct a detailed analysis of the understanding and generation parts separately.

4 Methodology

4.1 Training-free Compression Strategies

Large language models, a cornerstone of Unified model architectures, have been widely observed to contain significant redundancy across both depth (Gromov et al., 2025) and width dimensions (Ma et al., 2023). We next investigate how such redundancy manifests within Unified models.

Layer Dropping for Depth Pruning Transformer based large language models are stacked by multiple layers and scaling the depth of layers serves an effective way to enhance the performance. However, the depth also reflect the redundancy. Following Gromov et al. (2025); He et al. (2024), we measure the layer-wise importance via:

$$S_l = \text{Cosine\_Sim}(\mathbf{x}_l, \mathbf{y}_l), \tag{3}$$

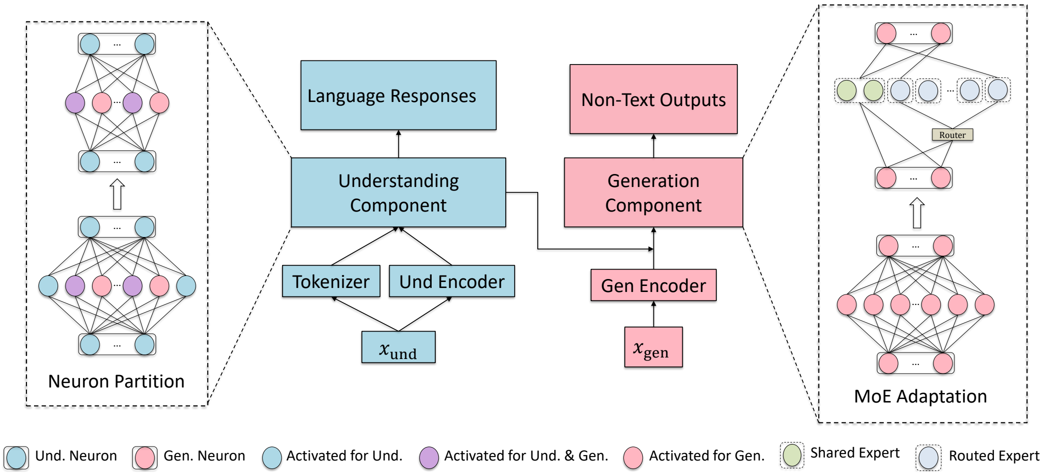


Figure 1: Overview of the proposed framework for unified multimodal model compression. The model is composed of an **understanding component** that processes multimodal inputs into embeddings and language responses, and a **generation component** that produces non-text outputs. We introduce two complementary strategies: **Neuron Partition**, which separates neurons into subsets and filter the neurons activated for the given task (Gen. in the figure); and **MoE Adaptation**, which dynamic activate neurons which have been partitioned into shared and routed experts managed by a router.

where $\mathbf{x}_l$ and $\mathbf{y}_l$ correspond to the input and output of the l-th layer, respectively. The similarity provides a measure of redundancy, with higher values implying that the layer contributes only marginal transformation. The metric has been shown to perform effectively in unimodal LLMs such as Mistral (Jiang et al., 2023) and LLaMA (Touvron et al., 2023b; Grattafiori et al., 2024). We next extend this evaluation to Unified models.

Width Reduction via Neuron Partition In addition to depth, scaling the width, particularly within MLP layers, has become a prevalent strategy for enhancing model capability. In general, an MLP layer expands the input from dimension d to dm through an up-projection and a gated projection, applies a nonlinear transformation, and then projects it back to dimension d via a down-projection. Here, m denotes the expansion multiplier, which increases hidden dimensionality to enhance model capacity but simultaneously introduces a substantial number of parameters. Given that MLP layers are expanded to dm hidden neurons, we further decompose them at the neuron level into important and less important counterparts.

To measure neuron importance, we draw inspiration from Wanda (Sun et al., 2024), which leverages both weights and activations as pruning metrics, and extend it from an unstructured to a structured neuron-level criterion. Given an input $x \in \mathbb{R}^{s\times d}$, in a Gate-Up-Down MLP, the hidden activations $h \in \mathbb{R}^{s\times dm}$ and output $y \in \mathbb{R}^{s\times d}$ can be written as:

$$h = \big(\text{SiLU}(xW_g^\top)\big) \odot (xW_u^\top), \qquad y = hW_d^\top, \tag{4}$$

where $W_g, W_u \in \mathbb{R}^{md\times d}$ are the up-projection matrix and gate-projection-matrix, $h \in \mathbb{R}^{n\times md}$ is the gated activation, $W_d \in \mathbb{R}^{d\times md}$ is the down-projection matrix. The hidden activations consist of md neurons, and the contribution of the i-th neuron to the final output is:

$$\Delta y_i = h_i W_{d,i}^\top, \tag{5}$$

with $W_{d,i}$ being the i-th column vector of W_d. If the i-th neuron is pruned, the induced output error norm can be approximated by:

$$\|\Delta y\|_2 \approx \|h_i W_{d,i}^\top\|_2. \tag{6}$$

Given all inputs from the calibration dataset $\mathcal{D}$, the accumulated error of each neuron is used as its importance metric:

$$s_i = \mathbb{E}_{x\sim\mathcal{D}}\big[|h_i| \cdot \|W_{d,i}^\top\|_2\big], \tag{7}$$

where $|h_i|$ measures the average activation magnitude of the neuron, and $\|W_{d,i}\|_2$ quantifies its amplification effect on the output. Therefore, neurons with larger scores play more critical roles, while those with smaller scores can be safely removed. Unlike unstructured pruning that zeroes individual weights, our approach enforces structured pruning by removing entire neurons, Concretely, this corresponds to removing column i from W_d and row i from both W_u and W_g, thereby ensuring hardware-friendly efficiency.

Unified models unify diverse tasks within a single architecture, and different tasks naturally activate different subsets of neurons. Figure 2 illustrates the distinct partitions activated by different tasks: the top 50% of important neurons identified from understanding and generation tasks overlap by only about 50%. This task-dependent variation reveals that redundancy is unevenly distributed: some neurons are indispensable for understanding but less relevant for generation, while others are critical for conditioning generative processes. To account for this heterogeneity, we apply the neuron-level importance metric across tasks to more accurately identify the principal neurons.

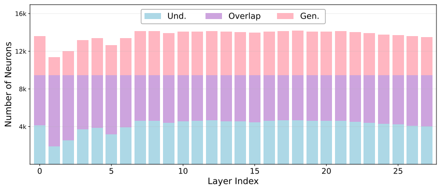


Figure 2: **Statistical analysis of high-importance neurons**, quantifying those predominantly activated in understanding tasks, in generation tasks, and jointly across both.

4.2 Training-Aware MoE Adaptation

Dynamic Activation Recognizing that the principal components vary across tasks, we next investigate activation patterns across different input samples. Figure 3 illustrates the activated neurons within a single layer of the generation component across multiple time steps (eight inputs, each with 30 denoising steps). This reveals a dynamic activation phenomenon, where the set of active parameters depends on the input, consistent with the intuition behind Mixture-of-Experts (MoE). To exploit this property, we integrate an MoE mechanism into Unified models through three key steps: Expert Partition, Expert-frozen Tuning, and MoE Adaptation.

Expert Partition To separate universal and task-specific capacity, we partition MLP neurons into *shared* and *routed* experts using cumulative importance across tasks. For each neuron i, let $s_i^{(t)}$ be its importance under task $t \in \mathcal{T}$. We compute the cumulative score:

$$S_i = \sum_{t \in \mathcal{T}} s_i^{(t)}. \tag{8}$$

The neurons with the highest S_i are selected as shared experts E_s, preserving features that consistently benefit multiple tasks (e.g., vision–language reasoning, image generation, or editing). The remaining neurons $\mathcal{R} = \{i \mid i \notin E_s\}$, which are more task-dependent, are evenly allocated to routed experts $\{E_r^{(1)}, \dots, E_r^{k}\}$ by ranked importance to ensure balanced capacity.

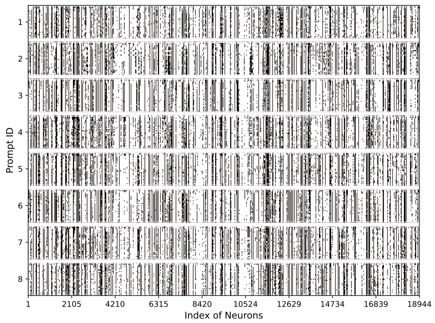


Figure 3: Visualization of dynamic activation patterns within a single layer of the generation component, evaluated on 8 input prompts over 30 denoising steps each.

MoE Adaptation After expert partition, we insert a router per layer to dynamically select routed experts for each input. In this case, the output of an MoE layer is formulated as follows:

$$\text{MoE}(x) = f_{\mathcal{S}}(x) + \sum_{j \in \text{Top-}k(\mathcal{G})} \mathcal{G}_j \cdot f_{\mathcal{R}_j}(x), \tag{9}$$

where $\mathcal{G}$ denotes the gating function, and $f_{\mathcal{S}}$ and $f_{\mathcal{R}}$ represent the transformations of shared and routed experts, respectively. The original MLP layer can be viewed as a special case of Equation 9, where all experts are selected. MoE adaptation adjusts the model to optimize performance with only a subset of activated parameters. To initialize this mechanism, we adopt a lightweight *expert-frozen tuning* stage as a cold start.

During expert-frozen tuning, the experts remain fixed and the remaining parameters are trainable. On the one hand, expert-frozen tuning leverages the capacity of existing experts without altering their pretrained knowledge. On the other hand, this enables the model to establish a preliminary routing policy, ensuring that experts acquire meaningful specialization before joint training. After this, we release the constraint of freezing experts to further optimize the performance.

5 Experiments

In this section, we present experiments on training-free compression and MoE adaptation for unified multimodal models.

5.1 Experimental Setup

Models We focus on several mainstream open-source Unified models, including BAGEL (Deng et al., 2025), Ming-Omni (AI et al., 2025), and Qwen-Image (Wu et al., 2025). All three adopt Qwen-Instruct (Yang et al., 2024) as the backbone for multimodal understanding. The key differences arise in their generation components: BAGEL employs a Mixture-of-Transformers (MoT) (Liang et al., 2025) design and reuses the Qwen-Instruct backbone for generation; Qwen-Image incorporates an MMDiT-based generator (Esser et al., 2024) and Ming-Omni adopts a multi-scale DiT block architecture. Table 1 presents a detailed comparison of these models.

Table 1: Summary of evaluated unified models.

Model	Und. Component	Und. Param.	Gen. Component	Gen. Param.
Qwen-Image	VLM	7.62B	MMDiT	20.42B
Ming-Omni	MLP	17.12B	MMDiT	2.51B
BAGEL	VLM	7.62B	LLM	7.62B

For flexible expert selection, each MoE layer is configured with 64 experts, including 8 shared experts following the design choices in Dai et al. (2024); DeepSeek-AI et al. (2025). The overall activation ratio is set to 50% per layer. All intermediate layers, except the first and the last, are converted into MoE layers.

Datasets For the calibration datasets used in training-free compression, we draw from multimodal understanding benchmarks (MME (Liu et al., 2023b), MMBench (Liu et al., 2023b), MMMU (Yue et al., 2023), MMVP (Tong et al., 2024)), image generation datasets (GenEval (Ghosh et al., 2023) and Wise (Niu et al., 2025)). For calibration in depth pruning or neuron partition, we use 128 training examples drawn from the same task type. For MoE adaptation, we additionally incorporate high-quality image–text pairs, complemented by a small amount of synthetic data generated by existing text-to-image models.

5.2 Understanding Components are Robust than Expected

Depth Reduction works in Generation Tasks but Fails in Understanding We begin by evaluating the impact of depth reduction. Since understanding components are less directly tied to image generation than generation components, we first examine this relatively less critical component and assess its effect on generation performance. Specifically, we remove transformer blocks, MLP layers, and attention layers, respectively. As shown in Figure 4, removing entire layers in the understanding component proves effective for BAGEL and Qwen-Image, but is less effective for Ming-Omni. We attribute this difference to architectural design: Ming-Omni's generation component is relatively smaller and thus depends more heavily on precise features encoded by the understanding component.

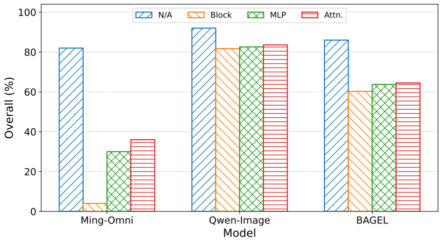


Figure 4: Comparison of the overall performance of depth reduction on the GenEval.

On the other hand, such compression substantially deteriorates the model's understanding capability. As shown in Table 6, removing half of the MLP layers causes performance on MME (Fu et al., 2023) to drop from 1684.8 to 304.5 in perception and from 696.7 to 127.1 in cognition. These results suggest that depth reduction fails to preserve the performance of the Unified model in both generation and understanding tasks. It is also worth noting that auto-regression is an error accumulation process, leading the model to collapse within only a few steps, as illustrated in Figure 9.

Table 2: Performance on GenEval when applying Neuron Partition to the understanding component. Since only the **understanding component** is compressed, the reported parameter counts correspond to this part rather than the full model size.

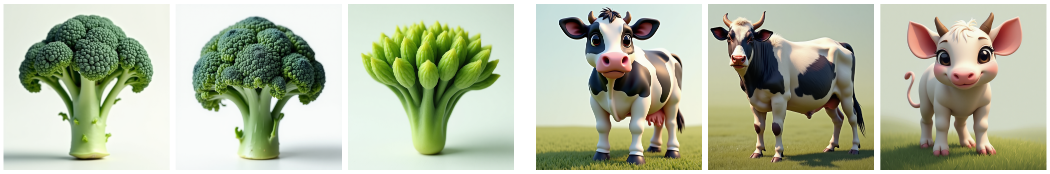
Model	Sparsity	Params.	Single Obj.	Two Obj.	Counting	Colors	Position‘	Color Attri.	Overall↑
BAGEL	0%	7.62B	0.99	0.94	0.81	0.95	0.72	0.77	0.86
	50%	4.76B	0.94	0.63	0.62	0.77	0.47	0.34	0.63
Qwen-Image	0%	7.62B	0.99	0.98	0.91	0.94	0.80	0.89	0.92
	50%	4.76B	0.99	0.94	0.94	0.93	0.76	0.87	0.90
	70%	3.62B	0.97	0.88	0.85	0.91	0.60	0.71	0.82
Ming-Omni	0%	17.12B	0.97	0.95	0.67	0.92	0.71	0.71	0.82
	50%	8.55B	0.97	0.92	0.66	0.89	0.61	0.70	0.79
	70%	5.61B	0.96	0.81	0.58	0.86	0.49	0.56	0.71

Neuron Partition on Understanding Components: Effective in Both Understanding and Generation In contrast to depth reduction, we propose Neuron Reduction, which prunes channels within the MLP layers. We first evaluate the effectiveness of this approach on the understanding components. Specifically, we compress the MLP layers to the target ratios (e.g., 50%) using a small set of calibration samples. As shown in Table 2, Ming-Omni and Qwen-Image largely maintain their performance even under aggressive compression ratios (i.e., 50% and 75%), whereas BAGEL exhibits a greater loss in capability, likely due to its mixture-of-transformers architecture (Liang et al., 2025), in which components interact more frequently through cross-attention at every layer. Similarly, neuron partition can be extended to attention heads, and it remains effective for compressing understanding components in generation tasks in Appendix A.

Similarly, understanding components are more compressible for neuron partition in generation tasks than in multimodal understanding. As shown in Table 3, neuron partition consistently achieves substantially better performance than depth reduction across all tasks. However, because understanding components directly affect the textual outputs in these tasks, their compression ratios should be kept more moderate than for generation tasks.

Table 3: Performance of neuron partition on understanding tasks.

Model	Sparse Ratio	MME-P	MME-C	MMMU	MMBench	MMVP
Ming-Omni	–	1584.3	670.4	66.7	86.73	54.6
	25%	1578.5	560.4	56.7	81.2	51.3
	50%	1269.0	317.9	51.7	81.0	46.0
BAGEL	–	1684.8	696.7	65.0	88.1	69.6
	25%	1558.1	681.7	60.1	85.7	68.7
	50%	916.5	276.1	56.7	79.21	56.0



(a) A realistic broccoli sits upright on a plain surface.

(d) A cow stands on a grassy field.

Figure 5: Impact of calibration data selection on multimodal generation. Each triplet shows outputs from the **unmodified model (left)**, the model after neuron partition with **image generation calibration (middle)**, and with **understanding calibration (right)**.

Calibration Data affects the Activated Parameters Neuron partition leverages calibration samples to estimate neuron importance and prunes those deemed less critical, as different tasks activate different subsets of neurons. To examine how the choice of calibration samples influences the retained parameters and the resulting performance, we conduct an ablation study using samples from understanding tasks (i.e., MME) and generation tasks (i.e., GenEval), respectively.

We find the alignment between calibration data and target tasks contributes to the performance. For instance, using samples from image generation would degrade the MMbench from 79.2 to 74.8. This trend also highlights in generation results shown in Figure 5. When calibrated with image generation samples (middle), the outputs remain faithful to the prompts, producing broccoli, scissors, skateboards, and cows with correct structures. In contrast, calibration with understanding samples (right) introduces distortions and mismatches.

This demonstrates that task-aligned calibration data yields better performance, while mismatched data degrades generation quality. The effect is particularly critical for unified models, where both input and output types vary in different combination of modalities.

5.3 Dilemma in Compressing Generation Component

We next investigate how compression influences generation quality by applying neuron partition or depth reduction to the generation components. While neuron partition yields promising efficiency gains, compressing the generation experts introduces a clear dilemma. As illustrated in Figure 6, aggressive compression severely compromises the fidelity and coherence of generated outputs. For instance, compressed models often produce distorted shapes and unrealistic textures, deviating from the intended semantics. This is consistent with observations from depth reduction and attention head reduction in Appendix C.

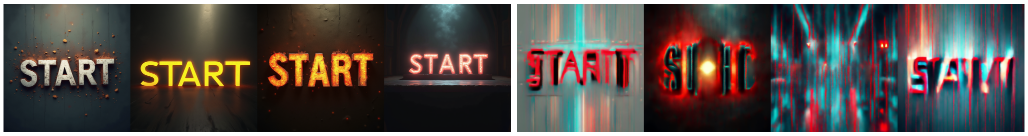


Figure 6: Qualitative comparison of baseline and compressed models. **The baseline model (left)** is tested without compression, while **the compressed model (right)** reduces the generator width by 50%. Results are shown for the prompts: "The word `START`" Compression leads to noticeable degradation in fine details and semantic consistency.

This highlights the contrasting compressibility between understanding and generation components: whereas understanding tasks remain robust under compression, generation quality is highly sensitive, limiting the extent of feasible compression.

5.4 MoE Adaptation

Given the potential performance degradation of compression (especially in the generation component) and the dynamic nature of principal activated components across tasks, static parameter partitioning fails to accurately capture the neurons required for activation. To address this limitation, we next explore MoE-based adaptation as a means to enhance performance.

Effectiveness of Expert-frozen Tuning After partitioning the experts, we first investigate the potential of existing experts by freezing them and training only the remaining parameters. This strategy mitigates catastrophic forgetting and encourages the model to learn effective expert selection while preserving pretrained knowledge. Specifically, we examine scenarios with different numbers of experts, comparing three configurations (16, 32, and 64) in Figure 7. The results show that finer-grained expert partitioning allows for more flexible activation combinations, leading to substantially lower training loss.

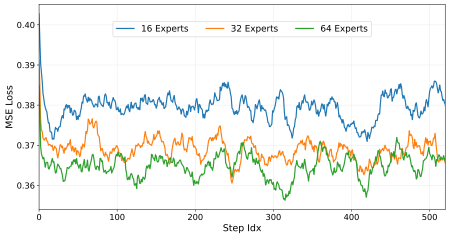


Figure 7: Expert-frozen training under different numbers of total experts.

Prior to tuning, the model produces noisy, low-detail images that fail to capture fine-grained semantics. With expert-fronzen tuning, however, we observe a steady decline in loss values in Figure 7, indicating stable convergence, and a substantial recovery in generation quality. For example, the overall GenEval score improves from 0.62 to 0.78, reflecting more coherent and visually faithful outputs. As illustrated in Figure 8, expert not only enhances image fidelity but also improves alignment between the generated content and the given instructions.

On the one hand, this demonstrates that certain subsets of parameters within the generation components, though difficult to compress, still retain the potential to produce high-quality images. On the other hand, adapting the routing mechanism alone can effectively unlock latent capacity within the experts, providing a lightweight yet powerful means of enhancing model performance under compression.

MoE Adaptation for Parameter Efficiency After the model learns to effectively select experts through Expert-fronzen Tuning, we release the constraint of frozen expert parameters to further enhance performance. Beyond applying MoE adaptation solely to the generation component, we also

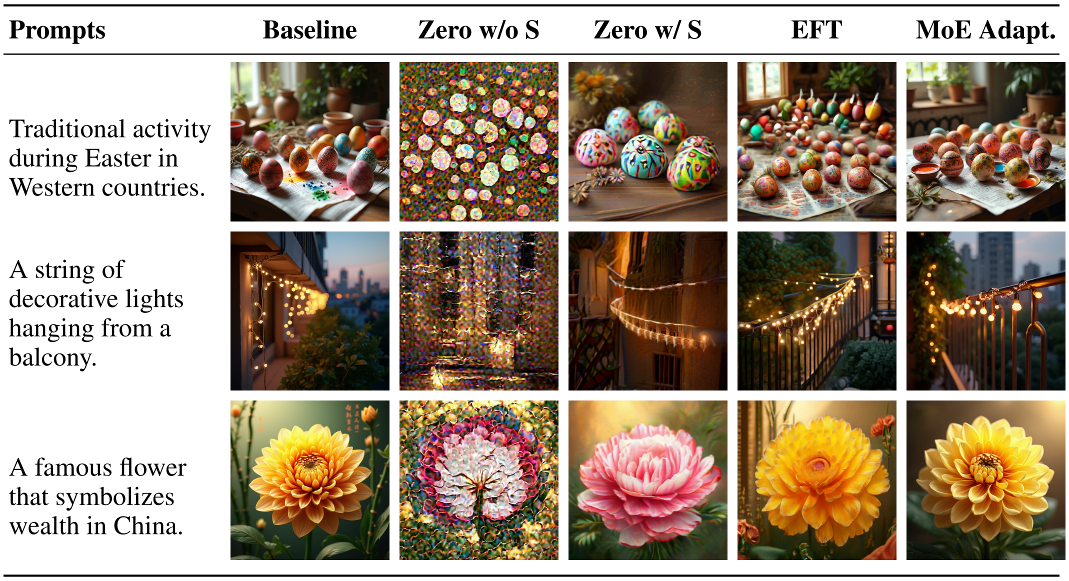

Prompts
Baseline
Zero w/o S
Zero w/ S
EFT
MoE Adapt.
Traditional activity during Easter in Western countries.
A string of decorative lights hanging from a balcony.
A famous flower that symbolizes wealth in China.

Figure 8: Comparison of models, including baseline without modification, zeroshot from expert partition with/without shared experts (Zero w/o S and Zero w/S), trained model after Expert-Frozen Tuning (EFT) and MoE Adapation (MoE Adapt.). The test prompts are sampled from WISE Niu et al. (2025).

explore transforming the understanding expert into an MoE structure, aiming to reduce the budget of activated parameters while preserving task effectiveness. To preserve fidelity on understanding tasks, we freeze the corresponding experts. Generation tasks, however, are more tolerant to sparsity, which enables us to apply sparse activation to the understanding experts for generation while keeping dense activation for understanding. In this case, we propose two versions of MoE adaptation: (1) applying expert partitioning and adaptation only to the generation experts, and (2) applying expert partitioning to both understanding and generation experts while keeping the understanding experts frozen and adapting generation experts only.

Unlike Expert-fronzen Tuning, which only updates the router while keeping experts frozen, MoE adaptation additionally enables training of the experts themselves. This extra training step allows the experts to refine their parameters based on the routing decisions, leading to better specialization and more accurate representations. As a result, in Table 4, the model achieves higher generation quality, demonstrating that fine-tuning both the router and experts is more effective than adjusting the router alone.

Table 4: Comparative performance across progressive stages of MoE adaptation, including Expert Partition without additional training, Expert-frozen Tuning, and full MoE Adaptation. For reference, results from the dense model with neuron partition under an equivalent budget of activated parameters are also reported.

Method	**Adapt. Comp.**	**Activated Params.**	**Single Obj.**	**Two Obj.**	**Counting**	**Colors**	**Position**	**Color Attri.**	**Overall↑**
Baseline	N/A	7.62B + 7.62B	0.99	0.94	0.81	0.95	0.72	0.77	0.86
Expert Partition	Gen.	7.42B + 4.96B	0.90	0.70	0.49	0.74	0.53	0.34	0.62
Dense Finetuning			0.97	0.88	0.75	0.91	0.67	0.71	0.82
Expert-frozen Tuning			0.99	0.94	0.62	0.93	0.69	0.54	0.78
MoE Adaptation			0.99	0.95	0.81	0.94	0.69	0.76	0.86
Expert Partition	Und & Gen	4.96B + 4.96B	0.69	0.18	0.23	0.45	0.10	0.05	0.28
Dense Finetuning			0.97	0.89	0.76	0.91	0.70	0.64	0.81
Expert-frozen Tuning			0.94	0.63	0.62	0.77	0.47	0.34	0.63
MoE Adaptation			0.99	0.96	0.78	0.95	0.70	0.72	0.85

6 Conclusion

Given the efficiency-oriented design of Omni-models that unify understanding and generation, we build on prior work in both training-free and training-aware compression. For training-free compression, we propose width reduction, demonstrating the high compressibility of understanding components when applied to generation tasks. Although compressing generation components presents greater challenges, our proposed MoE adaptation substantially recovers performance, enabling the trained model to match that of fully activated models. Together, these findings in training-free and training-aware compression offer valuable insights for the multimodal community.

Ethics Statement

Our work focuses on developing efficient architectures and compression methods for multimodal Omni-models. The techniques proposed are general-purpose and model-centric, without involving sensitive or personally identifiable information. We intend the released code and models to be used strictly for research and educational purposes, and will provide appropriate licensing terms to discourage potential misuse in harmful applications such as surveillance, disinformation, or other privacy-intrusive scenarios.

Reproducibility Statement

We ensure reproducibility through comprehensive documentation and code release. Specifically, we will provide: (1) source code implementing our pruning and MoE adaptation methods; (2) scripts and configuration files for replicating all main experiments; (3) fixed random seeds and hyperparameter settings; and (4) clear instructions for environment setup and evaluation.

Ablation studies and multiple-seed experiments reported in the paper further demonstrate the robustness and reproducibility of our findings.

References

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-omni: A unified multimodal model for perception and generation, 2025. URL `https://arxiv.org/abs/2506.09344`.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL `https://arxiv.org/abs/2303.12712`.

Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. URL `https://arxiv.org/abs/2401.06066`.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun

Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL `https://arxiv.org/abs/2412.19437`.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL `https://arxiv.org/abs/2505.14683`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. URL `https://openreview.net/forum?id=FPnUhsQJ5B`.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. URL `https://api.semanticscholar.org/CorpusID:259243928`.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=Wbr51vK331`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan

Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ngmEcEer8a.

Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. What matters in transformers? not all attention is needed, 2024. URL https://arxiv.org/abs/2406.15786.

Shwai He, Ang Li, and Tianlong Chen. Rethinking pruning for vision-language models: Strategies for effective sparsity. *SIGMETRICS Perform. Eval. Rev.*, 53(2):9–14, August 2025. ISSN 0163-5999. doi: 10.1145/3764944.3764948. URL https://doi.org/10.1145/3764944.3764948.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Nu6N69i8SB.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=w0H2xGHlkw.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281, 2023b. URL https://api.semanticscholar.org/CorpusID:259837088.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HtqnVSCj3q.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, 2023.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Chaoran Feng, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PxoFut3dWW.

Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. ECoFLap: Efficient coarse-to-fine layer-wise pruning for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=iIT02bAKzv.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024.

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

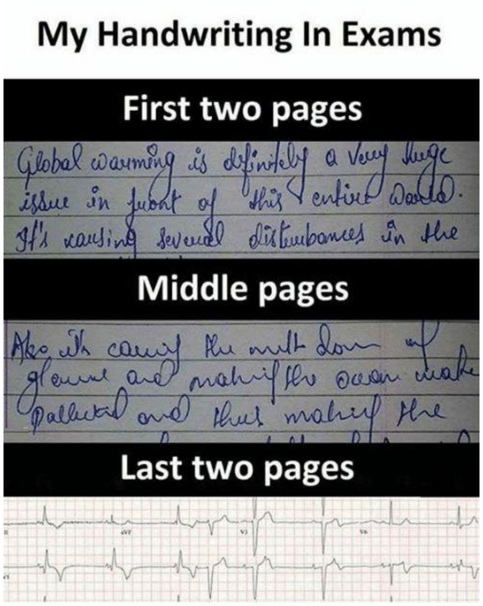
Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2023. URL https://api.semanticscholar.org/CorpusID:265466525.


My Handwriting In Exams
First two pages
Global warming is definitely a very huge issue in front of this entire world. It's causing several disturbances in the
Middle pages
Last two pages

Meme: “My Handwriting In Exams”

Question: Can someone explain what’s funny about this meme?

Neuron Partition:
The humor in this meme comes from the exaggerated change in handwriting style as the exam progresses. At the beginning, the handwriting is clear and legible, indicating a confident and well-prepared student. However, as the exam continues, the handwriting becomes increasingly difficult to read, suggesting that the student is becoming more anxious and less focused.

Depth Reduction:
The comedic resurgence biting meme portrays a comedic portrayal portraying a trope portraying idioms portraying idioms portraying idioms portraying idioms portraying idioms portraying idioms portraying exposes exposing exposing injuring injuring injuring injuring injuring injuring injuring injuring injuring injuring......

Figure 9: Comparison of neuron partition and depth reduction on a toy example under a 50% compression ratio. The model with fewer neurons per layer still answers the question reasonably well, whereas the model with reduced depth collapses and degenerates into repeating a single word.

A Attention Head Partition

While our main focus is width reduction in MLP layers, since they account for most of the parameters, our compression metric can be seamlessly extended to attention by computing the importance scores of attention heads. Attention Head Parttion still works for understanding components as shown in Table 5.

Table 5: Performance of attention head partition at a sparsity ratio of 50% per layer.

Model	Compressed Layers	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
BAGEL	N/A	0.99	0.94	0.81	0.95	0.72	0.77	0.86
	3-27	0.97	0.87	0.66	0.88	0.33	0.31	0.67
	4-27	0.98	0.91	0.72	0.89	0.41	0.40	0.72

B Dilemma of Depth Reduction on Understanding Tasks

While depth reduction has limited impact on generation tasks when applied to the understanding component, it fails on multimodal understanding tasks. Figure 9 shows that the reduced-depth model cannot generate continuous tokens in the answer. Nevertheless, the initial tokens remain reasonable, consistent with the role of the understanding component in generation tasks, which primarily performs prefilling and provides embeddings rather than full autoregressive decoding.

Table 6: Performance of depth reduction on understanding tasks.

Model	Sparsity	MME-P	MME-C	MMMU	MMBench	MMVP
Ming-Omni	–	1584.3	670.4	66.7	86.7	54.6
	50%	1197.2	308.2	51.7	81.2	46.0
BAGEL	–	1684.8	696.7	65.0	88.1	69.6
	50%	304.5	127.1	16.7	18.6	23.1

C More Results of Compressing Generation Component

Generation components are more sensitive to compression than understanding components. In addition to the results in Figure 6, we conduct experiments with depth reduction (Figure 10) and find that removing entire layers has a catastrophic effect on the output images. This suggests that preserving depth while compressing in width is a more effective strategy.

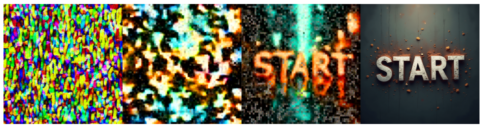
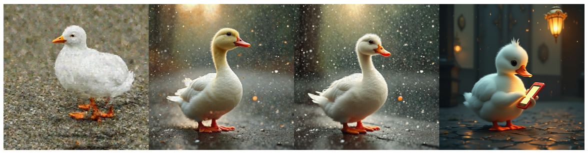


Figure 10: Depth reduction applied to MLP layers in the generation component. Figures are shown with decreasing numbers of removed layers: 14 (50%), 7 (25%), 4 (14%), and 0.

On the other hand, compressing the attention layers leads to substantial degradation in both depth and width settings. As shown in Figure 11, applying more than a 10% reduction results in noticeable performance drops.



Depth reduction achieved by removing 7, 4, 2, or 0 layers.

Figure 11: Compression of generation components through pruning of attention layers and heads.