MIXTURE-OF-MAMBA: ENHANCING MULTI-MODAL STATE-SPACE MODELS WITH MODALITY-AWARE SPARSITY

Weixin Liang*
Stanford University
wxliang@cs.stanford.edu

Junhong Shen*
Carnegie Mellon University
junhongs@andrew.cmu.edu

Genghan Zhang
Stanford University
zgh23@stanford.edu

Ning Dong
FAIR at Meta
dnn@meta.com

Luke Zettlemoyer
FAIR at Meta
lsz@meta.com

Lili Yu
FAIR at Meta
liliyu@meta.com

ABSTRACT

State Space Models (SSMs) have emerged as efficient alternatives to Transformers for sequential modeling, but their inability to leverage modality-specific features limits their performance in multi-modal pretraining. Here, we propose **Mixture-of-Mamba**, a novel SSM architecture that introduces *modality-aware sparsity* through modality-specific parameterization of the Mamba block. Building on Mixture-of-Transformers (W. Liang et al.), we extend the benefits of modality-aware sparsity to SSMs while preserving their computational efficiency. We evaluate Mixture-of-Mamba across three multi-modal pretraining settings: **Transfusion** (interleaved text and continuous image tokens with diffusion loss), **Chameleon** (interleaved text and discrete image tokens), and an extended three-modality framework incorporating **speech**. Mixture-of-Mamba consistently reaches the same loss values at earlier training steps with significantly reduced computational costs. In the Transfusion setting, Mixture-of-Mamba achieves equivalent image loss using only **34.76%** of the training FLOPs at the **1.4B** scale. In the Chameleon setting, Mixture-of-Mamba reaches similar image loss with just **42.50%** of the FLOPs at the **1.4B** scale, and similar text loss with just **65.40%** of the FLOPs. In the three-modality setting, MoM matches speech loss at **24.80%** of the FLOPs at the **1.4B** scale. Our ablation study highlights the synergistic effects of decoupling projection components, where joint decoupling yields greater gains than individual modifications. These results establish *modality-aware sparsity* as a versatile and effective design principle, extending its impact from Transformers to SSMs, setting new benchmarks in multi-modal pretraining.

1 INTRODUCTION

State Space Models (SSMs) (Gu et al., 2021; Gu & Dao, 2023) have emerged as efficient alternatives to Transformers for sequential modeling, offering linear scaling in sequence length and strong performance in single-modality tasks. Mamba, a recent SSM variant, has demonstrated exceptional efficiency and scalability across diverse tasks by leveraging advanced gating mechanisms and selective state-space scanning (Gu & Dao, 2023). Despite these advantages, SSMs, including Mamba, remain inherently dense, applying the same set of parameters across all input tokens, regardless of modality. This uniform parameterization limits their ability to capture modality-specific features, leading to suboptimal performance in multi-modal pretraining.

Recent efforts have extended SSMs to multi-modal tasks. Works like VLMamba (Qiao et al., 2024) and Cobra (Zhao et al., 2024) augment Mamba for vision-language modeling by adding LLaVA-style projection modules that map image features into the token space of Mamba. In the vision domain, Vision Mamba (Zhu et al., 2024) and VMamba (Liu et al., 2024c) incorporate bidirectional

*Equal contribution.

scanning schemes and selective 2D scanning paths for image patch modeling. Similarly, Mamba has been explored for diffusion-based image and video generation, as seen in DiffuSSM (Yan et al., 2024) and Zigma (Hu et al., 2024), which employ unique state-space scanning patterns. While these approaches demonstrate the adaptability of Mamba, they are orthogonal to our focus, which introduces **modality-aware sparsity** directly into the Mamba block itself.

A promising approach to address such limitations is **model sparsity**, exemplified by Mixture-of-Experts (MoE) (Jacobs et al., 1991; Eigen et al., 2013; Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Jiang et al., 2024; Sukhbaatar et al., 2024). MoE reduces computational load by activating only a subset of model components for each input token, allowing experts to specialize in specific aspects of the data. Despite its potential, MoE-based architectures face challenges such as imbalanced expert utilization, bi-level optimization instability, and inefficient load balancing (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Shen & Yang, 2021; Xu et al., 2024). These issues motivate the need for alternative sparse architectures that are computationally efficient and easier to optimize.

In multi-modal contexts, prior work (Bao et al., 2022b; Wang et al., 2022; Shen et al., 2023b; Lin et al., 2024) has introduced *modality-aware sparsity* in Transformer-based MoE architectures. These approaches activate specific experts or parameters based on modality, enabling models to specialize in handling diverse data types. Other methods fine-tune modality-specific modules atop dense LLM backbones (Wang et al., 2023; He et al., 2024; Shen et al., 2023a; 2024b). Such methods show that simple rule-based modality routing often outperforms learned routing, likely due to improved training stability and reduced optimization challenges.

The closest work to our approach is MoE-Mamba (Pióro et al., 2024) and the related Blackmamba architecture (Anthony et al., 2024), which interleave Mamba blocks with MoE-augmented MLP layers. While effective, these hybrid designs apply sparsity only to the MLP layers, leaving the dense Mamba blocks unmodified. In contrast, we present **Mixture-of-Mamba**, a novel architecture that directly introduces *modality-aware sparsity* into the Mamba block itself. Inspired by Mixture-of-Transformers (Liang et al., 2024), our approach dynamically selects modality-specific weights in every input processing component of Mamba, enabling stable and efficient multi-modal pretraining. Furthermore, prior work (Liang et al., 2024) shows that MoE techniques can complement sparse architectures like Mixture-of-Transformers, suggesting that Mixture-of-Mamba and MoE-based MLP sparsification can be combined to achieve further gains.

To rigorously evaluate Mixture-of-Mamba, we conduct experiments across three multi-modal pretraining settings:

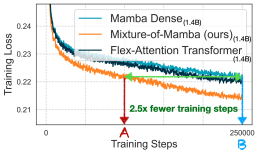
- **Transfusion:** Interleaved text and continuous image tokens with distinct autoregressive and diffusion-based objectives. Mixture-of-Mamba achieves equivalent image loss using only **34.76%** of the training FLOPs at the **1.4B** scale.

- **Chameleon:** Interleaved text and discrete image tokens. Mixture-of-Mamba reaches similar image loss with just **42.50%** of the FLOPs and similar text loss with only **65.40%** of the FLOPs at the **1.4B** scale.

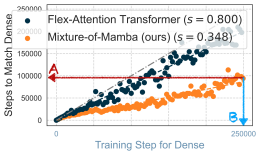
- **Three-Modality:** Extension of the Chameleon setting to include speech. Mixture-of-Mamba matches speech loss using only **24.80%** of the FLOPs at the **1.4B** scale, while maintaining strong performance across image and text modalities.

Additionally, we perform an ablation study to analyze the contribution of modality-specific parameterization. Our findings reveal a synergistic effect: jointly decoupling all components yields greater gains than individual modifications, underscoring the importance of modality-aware sparsity as a holistic design principle.

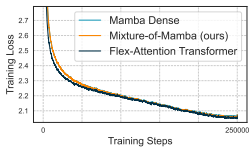
In summary, Mixture-of-Mamba establishes a versatile and efficient architecture for SSMs by extending *modality-aware sparsity* into the Mamba block. This approach delivers robust performance gains and substantial computational savings across diverse multi-modal settings, setting new benchmarks in scalable multi-modal pretraining. Code is released in https://github.com/Weixin-Liang/Mixture-of-Mamba.



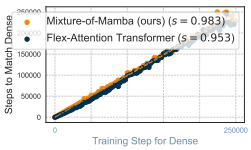
(a) **1.4B** Image Training Loss



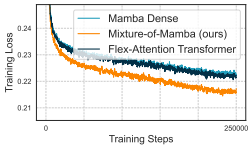
(b) **1.4B** Image Loss Matching



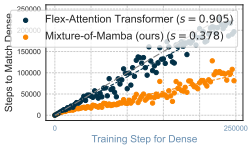
(c) **1.4B** Text Training Loss



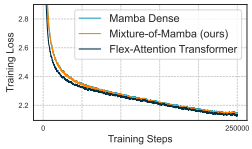
(d) **1.4B** Text Loss Matching



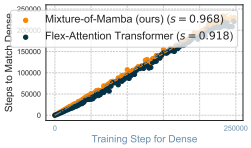
(e) **760M** Image Training Loss



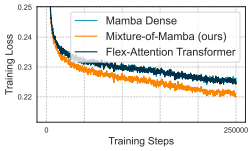
(f) **760M** Image Loss Matching



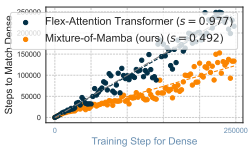
(g) **760M** Text Training Loss



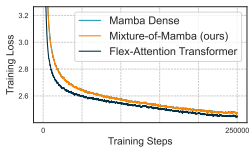
(h) **760M** Text Loss Matching



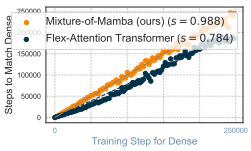
(i) **163M** Image Training Loss



(j) **163M** Image Loss Matching



(k) **163M** Text Training Loss



(l) **163M** Text Loss Matching

Figure 1: **Multi-modal pretraining in the Transfusion setting on interleaved text and image data across model scales.** Training loss and loss matching are reported for image and text modalities at three model sizes: **1.4B**, **760M**, and **163M**. **(a, e, i)** Image training loss shows significant improvements for Mixture-of-Mamba (orange), which consistently achieves lower loss compared to Mamba Dense (cyan) and Flex-Attention Transformer (dark gray) across all scales. **(b, f, j)** Image loss matching compares the training dynamics and shows that Mixture-of-Mamba and Flex-Attention Transformer reach the same loss values at earlier training steps compared to Mamba Dense. **(c, g, k)** Text training loss shows competitive results, with Mixture-of-Mamba performing better than Mamba Dense and on par with the Flex-Attention Transformer. **(d, h, l)** Text loss matching illustrates that Mixture-of-Mamba and Flex-Attention Transformer exhibit more efficient training dynamics than Mamba Dense, requiring fewer steps to achieve comparable loss values, though the primary improvements are observed in the image modality. Overall, in the **Transfusion setting**, Mixture-of-Mamba demonstrates substantial gains in image loss and training efficiency while maintaining strong performance on text.

2 Mixture-of-Mamba for Efficient Multi-Modal LLM Pretraining

2.1 Modality-Aware Sparsity in Mamba

The key novelty of Mixture-of-Mamba lies in integrating *modality-aware sparsity* directly into the Mamba block. By dynamically selecting modality-specific parameters for each input token based on its modality, our approach enables Mamba to efficiently process interleaved multi-modal sequences (e.g., text and image tokens) while preserving computational efficiency.

For interleaved multi-modal tokens $\{x_1, x_2, \ldots, x_T\}$ from multiple modalities, such as text and image, modality-specific parameterization dynamically selects the appropriate parameters for each token during processing. This general approach can apply to a wide range of transformations, such as linear, convolutional, and activation-based transformations. In Mamba, which primarily relies on linear transformations, the approach takes the following form:

$$f = Wx \quad \text{becomes} \quad f = \begin{cases} W_{\text{image}}x & \text{if } x \text{ is an image token} \\ W_{\text{text}}x & \text{if } x \text{ is a text token} \\ W_{\text{speech}}x & \text{if } x \text{ is a speech token} \end{cases}$$

Here, $W_{\text{image}}, W_{\text{text}}$, and W_{speech} are the modality-specific parameter matrices dynamically selected based on the modality of each token. While Mamba focuses on linear projections, the general technique of modality-aware sparsity can extend to other types of parameterized layers.

2.1.1 THE MIXTURE-OF-MAMBA BLOCK

The Mixture-of-Mamba block (Algorithm 2) builds on Mamba by dynamically applying modality-specific parameterization to key projections during input processing. This technique allows the block to handle interleaved multi-modal tokens more efficiently by leveraging modality-aware sparsity.

Each Mixture-of-Mamba block consists of input projection $W_{\text{in_proj}}$, intermediate projections $W_{\text{x_proj}}$ and $W_{\text{dt_proj}}$, and output projection $W_{\text{out_proj}}$, all parameterized by the token's modality using the general parameterization function $\mathcal{M}(X, W, b; M)$. The general form of the parameterization is given by:

$$\mathcal{M}(X, W, b; M) = \bigcup_{m \in M} \{X_m W_m + b_m\}$$

where X_m denotes the subset of tokens belonging to modality m, and W_m and b_m are the modality-specific parameters for that subset. This dynamic selection is applied at every stage of processing.

In Mixture-of-Mamba, projections explicitly processing input features belonging to a single modality—such as $W_{\text{in_proj}}$, $W_{\text{x_proj}}$, and $W_{\text{out_proj}}$—are decoupled using modality-specific parameters. However, components like Conv1D and state transitions A remain shared, as they operate across aggregated multi-modal features or RNN-like states, where the notion of modality is less well-defined. This design ensures computational efficiency while retaining modality-specific specialization.

Algorithm 1 Mixture-of-Mamba block

input $F_{in}, A, W_{in_proj}, W_{x_proj}, W_{dt_proj}, W_{out_proj}, b, M$
output F_{out}
$x, z \leftarrow \mathcal{M}(F_{in}, W_{in_proj}; M)$ ▷ Block starts
$u \leftarrow \text{SiLU}(\text{Conv1D}(x))$ ▷ [b,ℓ,d]
$\delta, B, C \leftarrow \mathcal{M}(u, W_{x_proj}; M)$ ▷ [b,ℓ,(r,n,n)]
$\Delta \leftarrow \log(1 + \exp((\mathcal{M}(\delta, W_{dt_proj}, b; M))))$
$\overline{A} \leftarrow \Delta * A$ ▷ [b,ℓ,d,n]
$\overline{B} \leftarrow \Delta * (u \times B)$ ▷ [b,ℓ,d,n]
$h = 0$ ▷ [b,d,n]
for $i = 0...N-1$ **do**
  $h = h * \overline{A_i} + \overline{B_i}$ ▷ [b,d,n]
  $y_i = h \cdot C_i$ ▷ [b,d]
end for
$o \leftarrow (y + u) * \text{SiLU}(z)$
$F_{out} \leftarrow \mathcal{M}(o, W_{out_proj}; M)$ ▷ Block ends

function $\mathcal{M}(X, W, b = \text{None}; M)$
  for each modality $m \in M$ **do**
    $I_m \leftarrow \{i : m_i = m\}$
    $X_m \leftarrow \{x_i : i \in I_m\}$
    $Y_m \leftarrow X_m W_m + b_m$
  end for
  return $Y \leftarrow \cup_{m \in M} Y_m$
end function

3 RESULTS

3.1 RESULTS IN MULTI-OBJECTIVE TRAINING (TRANSFUSION)

We evaluate Mixture-of-Mamba (MoM) against Mamba Dense and Flex-Attention Transformer in the **Transfusion setting**, where pretraining is performed on interleaved text and image data across three model scales: **163M**, **760M**, and **1.4B**. See our training configuration in Appendix Table 5. For clarity, performance gain is quantified as:

$$\text{Performance Gain (\%)} = \frac{\text{Loss}_{\text{Dense}} - \text{Loss}_{\text{Mixture}}}{\text{Loss}_{\text{Dense}}} \times 100,$$

where $\text{Loss}_{\text{Dense}}$ and $\text{Loss}_{\text{Mixture}}$ are the final losses of Mamba Dense and Mixture-of-Mamba, respectively. Relative training FLOPs reflect the computational cost required for MoM to match the training dynamics (similar loss) of Mamba Dense. The detailed results are summarized in Table 4 and Figure 1, with further visualizations provided in Appendix Figures 3, 4, and 5.

Image Modality. Mixture-of-Mamba (MoM) consistently demonstrates superior performance in **image modality training loss** across all model scales. At the **1.4B** scale, MoM achieves a training loss of **0.2138**, outperforming Mamba Dense by **2.20%** while requiring only **34.76%** of the training FLOPs. Similar trends are observed at smaller scales: at the **760M** scale, MoM achieves a training loss of **0.2172**, a **2.37%** improvement over Mamba Dense, while reducing training FLOPs to **37.76%**.

The validation loss curves on the CC12M dataset ((Table 4, Appendix Figure 4) further illustrate these trends. Mixture-of-Mamba consistently achieves lower image validation loss compared to Mamba Dense and Flex-Attention Transformer, with the improvements becoming more pronounced as model size increases. Additionally, loss matching curves demonstrate that MoM reaches equivalent loss values at earlier training steps, highlighting its improved training efficiency.

Text Modality. In the text modality, Mixture-of-Mamba consistently outperforms Mamba Dense across both training and validation metrics. At the **1.4B** scale, MoM achieves lower validation losses on both the C4 (**2.2695**) and Wikipedia (**1.7164**) datasets compared to Mamba Dense, despite their similar training losses. This indicates better generalization to unseen text data. Importantly, MoM also performs comparably to or better than Flex-Attention Transformer, particularly on validation losses, as shown in Appendix Figure 3. Similar trends are observed at smaller scales (**760M** and **163M**), where MoM reduces validation losses while maintaining high training efficiency.

Loss matching results in Appendix Figure 3 (b, f, j) confirm that Mixture-of-Mamba aligns closely with or surpasses Mamba Dense, reaching comparable loss values earlier during training. These improvements highlight MoM's strong performance in text tasks while maintaining its computational efficiency.

Overall Performance and Efficiency. Across both image and text modalities, Mixture-of-Mamba consistently outperforms Mamba Dense in terms of loss reduction while requiring significantly fewer training FLOPs to achieve similar learning dynamics. At the **1.4B** scale, MoM improves the overall training loss by **0.84%** while requiring only **83.10%** of the training FLOPs. At smaller scales, such as **760M** and **163M**, MoM reduces the overall training loss by up to **0.94%**, while requiring just **82.94%** and **86.11%** of the FLOPs, respectively (Table 4, Appendix Figure 5). These results, summarized in Table 4 and Figure 1, and further supported by Appendix Figures 3, 4, and 5, underscoring MoM's effectiveness, scalability, and efficiency in the **Transfusion setting**.

4 CONCLUSION

In this work, we introduced **Mixture-of-Mamba**, a novel extension of state-space models (SSMs) that incorporates *modality-aware sparsity* through modality-specific parameterization. By enabling modality-specific specialization while preserving the computational efficiency of SSMs, Mixture-of-Mamba consistently outperforms dense baselines across three multi-modal settings: Transfusion (interleaved text and continuous image tokens), Chameleon (interleaved text and discrete image

tokens), and an extended Chameleon+Speech framework. Our results demonstrate substantial improvements in loss reduction, with training efficiency gains reaching more than **double** the computational efficiency compared to dense SSMs. Ablation studies further reveal a synergistic effect from jointly decoupling key projection components, highlighting the effectiveness of modality-aware sparsity. These findings establish Mixture-of-Mamba as a scalable and efficient architecture for multi-modal pretraining, paving the way for future exploration in dynamic sparsity and broader multi-modal applications.

REFERENCES

Quentin Anthony, Yury Tokpanov, Paolo Glorioso, and Beren Millidge. Blackmamba: Mixture of experts for state-space models. *arXiv preprint arXiv:2402.01771*, 2024.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022a.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, 2022b. URL https://arxiv.org/abs/2111.02358.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. URL https://arxiv.org/abs/2405.09818.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. URL https://arxiv.org/abs/2101.03961.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Scalable diffusion models with state space backbone. *arXiv preprint arXiv:2402.05608*, 2024.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Wanggui He, Siming Fu, Mushui Liu, Xierui Wang, Wenyi Xiao, Fangxun Shu, Yi Wang, Lei Zhang, Zhelun Yu, Haoyuan Li, et al. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. *arXiv preprint arXiv:2407.07614*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Schusterbauer, and Björn Ommer. Zigma: A dit-style zigzag mamba diffusion model. In *ECCV*, 2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

P. Langley. Crafting papers on machine learning. In Pat Langley (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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, 2020. URL https://arxiv.org/abs/2006.16668.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

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

Alexander H. Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and James R. Glass. Dinosr: Self-distillation and online clustering for self-supervised speech representation learning, 2024a. URL https://arxiv.org/abs/2305.10005.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models, 2024b. URL https://arxiv.org/abs/2409.10695.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c. URL https://openreview.net/forum?id=ZgtLQQR1K7.

Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon Camarasa. Multiway-adapater: Adapting large-scale multi-modal models for scalable image-text retrieval. *arXiv preprint arXiv:2309.01516*, 2023.

Shentong Mo and Yapeng Tian. Scaling diffusion mamba with bidirectional ssms for efficient image and video generation. *arXiv preprint arXiv:2405.15881*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Maciej Pióro, Kamil Ciebiera, Krystian Król, Jan Ludziejewski, and Sebastian Jaszczur. Moe-mamba: Efficient selective state space models with mixture of experts, 2024.

Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538, 2017. URL http://arxiv.org/abs/1701.06538.

Junhong Shen and Lin F. Yang. Theoretically principled deep rl acceleration via nearest neighbor function approximation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11): 9558–9566, May 2021. doi: 10.1609/aaai.v35i11.17151. URL https://ojs.aaai.org/index.php/AAAI/article/view/17151.

Junhong Shen, Abdul Hannan Faruqi, Yifan Jiang, and Nima Maftoon. Mathematical reconstruction of patient-specific vascular networks based on clinical images and global optimization. *IEEE Access*, 9:20648–20661, 2021. doi: 10.1109/ACCESS.2021.3052501.

Junhong Shen, Mikhail Khodak, and Ameet Talwalkar. Efficient architecture search for diverse tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Junhong Shen, Liam Li, Lucio M. Dery, Corey Staten, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Cross-modal fine-tuning: align then refine. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Junhong Shen, Atishay Jain, Zedian Xiao, Ishan Amlekar, Mouad Hadji, Aaron Podolny, and Ameet Talwalkar. Scribeagent: Towards specialized web agents using production-scale workflow data, 2024a. URL https://arxiv.org/abs/2411.15004.

Junhong Shen, Tanya Marwah, and Ameet Talwalkar. Ups: Towards foundation models for pde solving via cross-modal adaptation. *arXiv preprint arXiv:2403.07187*, 2024b.

Junhong Shen, Neil Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolo Fusi. Tag-llm: Repurposing general-purpose llms for specialized domains, 2024c.

Junhong Shen, Kushal Tirumala, Michihiro Yasunaga, Ishan Misra, Luke Zettlemoyer, Lili Yu, and Chunting Zhou. Cat: Content-adaptive image tokenization, 2025. URL https://arxiv.org/abs/2501.03120.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023b.

Yikang Shen, Zhen Guo, Tianle Cai, and Zengyi Qin. Jetmoe: Reaching llama2 performance with 0.1 m dollars. *arXiv preprint arXiv:2404.07413*, 2024d.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen tau Yih, Jason Weston, and Xian Li. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm, 2024. URL https://arxiv.org/abs/2403.07816.

Renbo Tu, Nicholas Roberts, Mikhail Khodak, Junhong Shen, Frederic Sala, and Ameet Talwalkar. NAS-bench-360: Benchmarking neural architecture search on diverse tasks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. URL https://arxiv.org/abs/2208.10442.

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. Specialized foundation models struggle to beat supervised baselines, 2024. URL https://arxiv.org/abs/2411.02796.

Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8239–8249, 2024.

Luyao Yuan, Zipeng Fu, Jingyue Shen, Lu Xu, Junhong Shen, and Song-Chun Zhu. Emergence of pragmatics from referential game between theory of mind agents, 2021a. URL https://arxiv.org/abs/2001.07752.

Luyao Yuan, Dongruo Zhou, Junhong Shen, Jingdong Gao, Jeffrey L Chen, Quanquan Gu, Ying Nian Wu, and Song-Chun Zhu. Iterative teacher-aware learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29231–29245. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f48c04ffab49ff0e5d1176244fdfb65c-Paper.pdf.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520*, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2024.

A Extended Method

A.1 Multi-objective Training with Diffusion

Following Transfusion Zhou et al. (2024), Mixture-of-Mamba is trained on interleaved multi-modal sequences of discrete text tokens and continuous image tokens using a combined objective that incorporates both language modeling and diffusion-based image generation. Each image is encoded as a sequence of latent patches using a Variational Autoencoder (VAE), where each patch is represented as a continuous vector. The patches are sequenced left-to-right, top-to-bottom, and inserted into the discrete text sequence.

The diffusion process follows the Denoising Diffusion Probabilistic Models (DDPM) Ho et al. (2020), where Gaussian noise is progressively added to the latent image patches during the forward process. Given a clean latent patch $\mathbf{x}_0$, a noised version $\mathbf{x}_t$ at timestep t is created as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t$ is determined by a cosine noise schedule Nichol & Dhariwal (2021), approximated as $\sqrt{\bar{\alpha}_t} \approx \cos(\frac{t}{T} \cdot \frac{\pi}{2})$ with adjustments. During training, noise is added to the latent patches at a randomly selected timestep t, and the model is optimized to predict the noise $\boldsymbol{\epsilon}$.

The overall training objective combines the autoregressive language modeling loss $\mathcal{L}_{\text{LM}}$, applied to the discrete text tokens, with the diffusion loss $\mathcal{L}_{\text{DDPM}}$, applied to the latent image patches:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{DDPM}}, \tag{2}$$

where λ balances the contributions of the two losses.

Importantly, the conditioning for image generation is naturally embedded within the interleaved sequence. When denoising image patches, the preceding tokens—including both text describing the image and prior images—serve as context for conditional generation. This unified approach enables Mixture-of-Mamba to leverage the modality-aware sparsity to efficiently model both local intra-image dependencies and long-range inter-modal relationships across the sequence.

A.2 Training with Uniform Representations

As an alternative to the multi-objective training paradigm, we explore a unified representation strategy in which both text and image modalities are represented as discrete tokens. Following the Chameleon framework Chameleon Team (2024), we treat the image data as sequences of discrete tokens obtained through a pre-trained VQ-VAE model Gafni et al. (2022). Specifically, each image is encoded into a fixed number of tokens (e.g., 1,024) by quantizing its latent features into a learned codebook. These tokens are then arranged sequentially, similar to the processing of text tokens, resulting in a uniform discrete representation across both modalities.

During training, both text and image tokens are processed using the same autoregressive objective, where the model learns to predict the next token in the sequence given all previous tokens. Formally, the training objective is:

$$\mathcal{L}_{\text{uniform}} = \mathbb{E}_{\mathbf{x}_{1:T}} \left[-\log P(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}) \right], \tag{3}$$

where $\mathbf{x}_{1:T}$ represents the interleaved sequence of text and image tokens. This objective allows the model to treat text and image data equivalently, unifying the training process across modalities while relying solely on an autoregressive loss. The use of discrete tokens for images simplifies the training procedure by removing the need for separate loss formulations, as in the diffusion-based approach. It also aligns with the inherent sequence-to-sequence nature of Mixture-of-Mamba, where the same modality-aware sparsity design can be applied seamlessly across the discrete text and image tokens.

Motivation and Robustness Testing. We include this alternative strategy to evaluate the robustness of our Mixture-of-Mamba architecture under different choices of training objectives and data representations. By experimenting with uniform discrete representations, we demonstrate that Mixture-of-Mamba consistently outperforms Mamba Dense models across various settings, including both continuous (multi-objective) and discrete (uniform) representations. This highlights the versatility of Mixture-of-Mamba and its ability to deliver performance gains regardless of the underlying choice of modality representations or training objectives.

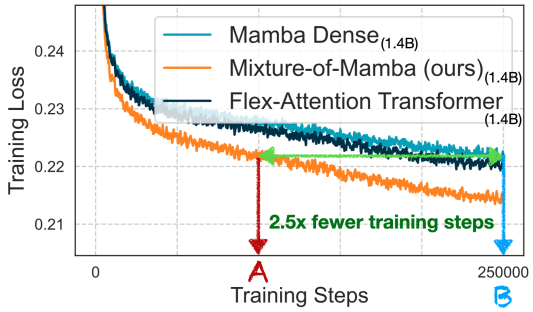


Figure 2: **Multi-modal pretraining on interleaved text and image data.** Training loss on the image modality is shown for models with 1.4B parameters: Mamba Dense (cyan), Flex-Attention Transformer (dark gray), and Mixture-of-Mamba (orange). The Mixture-of-Mamba achieves significantly lower training loss and requires **2.5x fewer training steps** (indicated by the green arrow) to reach the same loss level as the other baselines.

Model Scale	Metric Category	Metric Name	Mamba Loss (↓)	Mixture-of-Mamba Loss (↓)	Performance Gain (%) (↑)	Relative Training FLOPs to Match Mamba (%) (↓)
443M	Image Metrics	Training Loss	5.3558	5.1703	3.46%	33.40%
		Obelisc Val. Loss	4.5258	4.3546	3.78%	35.10%
		SSTK Val. Loss	5.9179	5.7471	2.89%	35.30%
	Text Metrics	Training Loss	2.4637	2.3864	3.14%	62.00%
		Obelisc Val. Loss	3.0544	2.9820	2.37%	66.70%
		SSTK Val. Loss	2.7569	2.6250	4.78%	54.70%
	Overall	Avg Training Loss	3.6584	3.5364	3.33%	47.90%
880M	Image Metrics	Training Loss	5.2260	5.1201	2.03%	48.40%
		Obelisc Val. Loss	4.4127	4.3105	2.32%	49.30%
		SSTK Val. Loss	5.7987	5.6986	1.73%	50.50%
	Text Metrics	Training Loss	2.3073	2.2438	2.75%	65.60%
		Obelisc Val. Loss	2.8886	2.8313	1.99%	72.80%
		SSTK Val. Loss	2.5483	2.4548	3.67%	67.90%
	Overall	Avg Training Loss	3.5130	3.4320	2.31%	58.30%
1.5B	Image Metrics	Training Loss	5.1892	5.0591	2.51%	42.50%
		Obelisc Val. Loss	4.3692	4.2510	2.71%	44.50%
		SSTK Val. Loss	5.7546	5.6335	2.10%	44.60%
	Text Metrics	Training Loss	2.2284	2.1614	3.01%	65.40%
		Obelisc Val. Loss	2.8020	2.7393	2.24%	71.60%
		SSTK Val. Loss	2.4614	2.3455	4.71%	62.10%
	Overall	Avg Training Loss	3.4602	3.3670	2.69%	54.70%

Table 1: **Training and validation metrics across model scales in the Chameleon setting.** In this setting, both image and text modalities are represented as discrete tokens. Mixture-of-Mamba achieves substantial performance improvements over Mamba Dense, with the **image modality** showing the largest gains. The **text modality** also exhibits significant improvements, in contrast to the Transfusion setting where text gains were more modest. The current table shows results for three model scales: **443M**, **880M**, and **1.5B**, due to space constraints. See Appendix Table 7 for the full results across all five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. These results further highlight the effectiveness and efficiency of Mixture-of-Mamba, which consistently achieves strong performance with reduced relative training FLOPs.

B Extended Results

B.1 Results in Training with Uniform Representations (Chameleon)

We evaluate Mixture-of-Mamba (MoM) in the **Chameleon setting**, where both **image** and **text** modalities are represented as discrete tokens. See our training configuration in Appendix Table 6. Results are summarized in Table 1, with full results across all five scales (**37M**, **94M**, **443M**, **880M**, and **1.5B**) provided in Appendix Table 7. Training dynamics and validation loss trends are visualized in Appendix Figures 6, 7, and 8.

Model Scale	Metric Category	Metric Name	Mamba Loss (↓)	Mixture-of-Mamba Loss (↓)	Performance Gain (%) (↑)	Relative Training FLOPs to Match Mamba (%) (↓)
37M	Speech Metrics	Training Loss	1.8159	1.6909	6.88%	10.30%
		LL60K Val. Loss	1.6756	1.5217	9.18%	13.60%
		PPL30K Val. Loss	1.8147	1.6845	7.17%	13.60%
	Overall Metrics	Avg Training Loss	4.2299	4.0759	3.64%	45.00%
94M	Speech Metrics	Training Loss	1.6911	1.5662	7.38%	11.90%
		LL60K Val. Loss	1.5235	1.3747	9.76%	14.80%
		PPL30K Val. Loss	1.6951	1.6152	4.71%	12.60%
	Overall Metrics	Avg Training Loss	3.7756	3.6371	3.67%	43.10%
443M	Speech Metrics	Training Loss	1.5414	1.4313	7.14%	19.20%
		LL60K Val. Loss	1.3466	1.2113	10.05%	24.70%
		PPL30K Val. Loss	1.5634	1.4790	5.40%	22.00%
	Overall Metrics	Avg Training Loss	3.3317	3.2096	3.66%	44.00%
880M	Speech Metrics	Training Loss	1.4902	1.4054	5.69%	22.40%
		LL60K Val. Loss	1.2939	1.1757	9.13%	30.10%
		PPL30K Val. Loss	1.5400	1.4619	5.07%	24.30%
	Overall Metrics	Avg Training Loss	3.2289	3.1571	2.22%	54.30%
1.5B	Speech Metrics	Training Loss	1.4790	1.3940	5.75%	24.80%
		LL60K Val. Loss	1.2592	1.1552	8.26%	32.10%
		PPL30K Val. Loss	1.5200	1.4387	5.35%	27.60%
	Overall Metrics	Avg Training Loss	3.1507	3.0545	3.05%	56.20%

Table 2: **Training and validation metrics across model scales with three modalities: image, text, and speech.** This setting extends the Chameleon framework by incorporating **speech** alongside image and text, with all modalities represented as discrete tokens. Mixture-of-Mamba achieves consistent improvements over Mamba Dense across all scales (**37M**, **94M**, **443M**, **880M**, and **1.5B**), particularly in the **speech modality**, where performance gains reach up to **9.18%**. These gains are achieved with substantial reductions in training FLOPs, ranging from **10.30%** to **56.20%** relative to Mamba Dense. The results demonstrate that Mixture-of-Mamba generalizes effectively to a multi-modal setting with three modalities while delivering significant computational efficiency.

Image Modality. Mixture-of-Mamba (MoM) consistently demonstrates better performance in **image modality training loss** across all model scales, achieving substantial efficiency gains over Mamba Dense. At the **443M** scale, MoM achieves a training loss of **5.1703**, a **3.46%** improvement over Mamba Dense, while requiring only **33.40%** of the training FLOPs. Similar trends are observed at other scales: at the largest **1.5B** scale, MoM achieves a training loss of **5.0591**, a **2.51%** improvement, with only **42.50%** of the training FLOPs. At the smallest **37M** scale, MoM reduces training loss to **5.9561**, outperforming Mamba Dense by **2.85%** while requiring just **25.90%** of the FLOPs (Appendix Table 7). These results highlight MoM's ability to achieve improved performance and convergence efficiency consistently in the image modality across all model scales.

Text Modality. Mixture-of-Mamba (MoM) demonstrates consistent improvements in **text modality training loss** across all model scales. At the largest **1.5B** scale, MoM reduces training loss to **2.1614**, a **3.01%** improvement over Mamba Dense, while requiring only **65.40%** of the training FLOPs. Validation loss on Obelisc and a proprietary version of the Shutterstock datasets (SSTK) exhibits similar trends, with MoM achieving notable improvements in loss values while maintaining significant efficiency gains (Appendix Figures 7 and 8). These results further highlight MoM's ability to deliver strong text performance with improved convergence efficiency. These results highlight Mixture-of-Mamba's robust and efficient improvements in the Chameleon setting across both image and text modalities, with substantial computational savings.

B.2 Results in Training with Three Modalities (Chameleon+Speech)

To evaluate the robustness and scalability of Mixture-of-Mamba (MoM), we extend the Chameleon framework to include a third modality: **speech**, alongside image and text, with all modalities represented as discrete tokens. Speech data is tokenized using an in-house tokenizer, a variant of DinoSR (Liu et al., 2024a), which extracts semantic tokens with a vocabulary size of 500, where each token corresponds to 40ms of audio content. Results are summarized in Table 2, with additional training dynamics and evaluation loss trends visualized in Appendix Figures 10, 11, 12, and 13.

Speech Modality. Mixture-of-Mamba (MoM) achieves substantial improvements in **speech modality training loss** across all model scales. At the **443M** scale, MoM improves speech training loss by **7.14%** compared to Mamba Dense. To match the training loss achieved by Mamba Dense, MoM requires only **19.20%** of the training FLOPs, demonstrating significant efficiency gains. Similar trends hold at the largest **1.5B** scale, where MoM achieves a **5.75%** improvement in speech training loss and matches Mamba Dense's loss with just **24.80%** of the training FLOPs.

Overall training loss is consistently reduced across scales. At the **1.5B** scale, MoM lowers the overall training loss by **3.05%**. When targeting the same loss as Mamba Dense, MoM achieves this with a **56.20%** reduction in relative training FLOPs, highlighting its improved computational efficiency.

Performance in the **image** and **text** modalities similarly shows consistent improvements in training and validation losses relative to Mamba Dense. Full results and trends are presented in Appendix Figures 12 and 13, where MoM's robust performance across all three modalities is further validated.

C Related Work

C.1 State-Space Models and Multi-Modal Extensions

State-space models (SSMs) (Gu et al., 2021; Gu & Dao, 2023) have recently gained traction as computationally efficient alternatives to Transformers for sequential modeling. Mamba (Gu & Dao, 2023), in particular, demonstrates strong performance on single-modality tasks by leveraging linear time complexity and advanced gating mechanisms. Extending Mamba to multi-modal tasks remains an active research area.

In vision-language modeling, VLMamba (Qiao et al., 2024) and Cobra (Zhao et al., 2024) augment Mamba by incorporating LLaVA-style projection modules, enabling image features to be mapped into the token space of the Mamba model for sequence modeling. In the vision domain, Vision Mamba (Zhu et al., 2024) introduces bidirectional scanning by chaining forward and backward SSM blocks, while VMamba (Liu et al., 2024c) further enhances image patch processing with a 2D Selective Scan (SS2D) module that traverses patches across multiple scanning paths.

For diffusion-based models, works such as DiffuSSM (Yan et al., 2024) and Zigma (Hu et al., 2024) replace attention mechanisms with SSMs for image and video generation. Zigma introduces a zigzag scanning scheme to improve efficiency for sequential diffusion tasks, while other approaches (Mo & Tian, 2024; Fei et al., 2024) explore bi-directional SSM architectures. While these works highlight the flexibility of Mamba in generative tasks, they focus primarily on architectural modifications for specific domains rather than general multi-modal pretraining.

The most related work to ours is MoE-Mamba (Pióro et al., 2024) and Blackmamba (Anthony et al., 2024), which interleave Mamba blocks with MoE-augmented MLPs to introduce sparsity. However, these hybrid designs apply sparsity only to the MLP layers, leaving the dense Mamba block unmodified. In contrast, our proposed Mixture-of-Mamba integrates modality-aware sparsity directly into the Mamba block by decoupling its projection components, enabling specialized computations for different modalities. This general design complements existing methods and offers new opportunities for computationally efficient multi-modal pretraining.

C.2 Sparse Architectures for Multi-Modal Pretraining

Model sparsity, particularly Mixture-of-Experts (MoE), has been extensively explored in Transformers to reduce computational cost (Jacobs et al., 1991; Eigen et al., 2013; Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Jiang et al., 2024). MoE selectively activates subsets of parameters for each input token, allowing the model to specialize in different aspects of the data. However, challenges such as expert imbalance, bi-level optimization, and load balancing remain prevalent (Shazeer et al., 2017; Lepikhin et al., 2020; Tu et al., 2022).

In multi-modal tasks, modality-aware sparsity has emerged as an effective strategy. Works such as VLMo (Shen et al., 2023b), MoMA (Lin et al., 2024), and related approaches (Wang et al., 2022; Shen et al., 2022; Bao et al., 2022a; Long et al., 2023; Shen et al., 2025) assign modality-specific

experts to handle the unique statistical properties of text, images, and other data types. This improves specialization while avoiding the complexities of learned routing mechanisms (Liang et al., 2022).

Transformer-based architectures have further extended sparsity into attention mechanisms (Shen et al., 2024d;c; Yuan et al., 2021b;a; Shen et al., 2021; Liu et al., 2024b; Shen et al., 2024a). CogVLM (Wang et al., 2023) applies sparse techniques on top of a pre-trained Vicuna-7B model but remains limited to generating text outputs. Concurrently, Playground v3 (PGv3) (Liu et al., 2024b) integrates DiT-style image transformers with a frozen LLaMA-3 backbone to achieve state-of-the-art performance in text-to-image generation.

Our work differs fundamentally in two key aspects. First, Mixture-of-Mamba introduces *modality-aware sparsity* into the Mamba block itself, generalizing sparse architectures beyond Transformers to SSMs. Unlike prior works that sparsify only the MLP or attention components, we decouple projection components of the Mamba block, enabling efficient and specialized computations across modalities. Second, Mixture-of-Mamba is trained from scratch for multi-modal generation tasks, unlike approaches like CogVLM and PGv3 that fine-tune pre-trained backbones.

Furthermore, our design is complementary to existing MoE techniques. Prior work (Liang et al., 2024) has demonstrated that MoE-based sparsification can be combined with sparse architectures like Mixture-of-Transformers to achieve additional gains. Similarly, Mixture-of-Mamba can serve as a versatile and computationally efficient solution, offering new pathways for scalable multi-modal pretraining.

D Extended Method

D.1 Multi-objective Training with Diffusion

Following Transfusion Zhou et al. (2024), Mixture-of-Mamba is trained on interleaved multi-modal sequences of discrete text tokens and continuous image tokens using a combined objective that incorporates both language modeling and diffusion-based image generation. Each image is encoded as a sequence of latent patches using a Variational Autoencoder (VAE), where each patch is represented as a continuous vector. The patches are sequenced left-to-right, top-to-bottom, and inserted into the discrete text sequence.

The diffusion process follows the Denoising Diffusion Probabilistic Models (DDPM) Ho et al. (2020), where Gaussian noise is progressively added to the latent image patches during the forward process. Given a clean latent patch $\mathbf{x}_0$, a noised version $\mathbf{x}_t$ at timestep t is created as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

where $\bar{\alpha}_t$ is determined by a cosine noise schedule Nichol & Dhariwal (2021), approximated as $\sqrt{\bar{\alpha}_t} \approx \cos(\frac{t}{T} \cdot \frac{\pi}{2})$ with adjustments. During training, noise is added to the latent patches at a randomly selected timestep t, and the model is optimized to predict the noise $\boldsymbol{\epsilon}$.

The overall training objective combines the autoregressive language modeling loss $\mathcal{L}_{\text{LM}}$, applied to the discrete text tokens, with the diffusion loss $\mathcal{L}_{\text{DDPM}}$, applied to the latent image patches:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{DDPM}}, \tag{5}$$

where λ balances the contributions of the two losses.

Importantly, the conditioning for image generation is naturally embedded within the interleaved sequence. When denoising image patches, the preceding tokens—including both text describing the image and prior images—serve as context for conditional generation. This unified approach enables Mixture-of-Mamba to leverage the modality-aware sparsity to efficiently model both local intra-image dependencies and long-range inter-modal relationships across the sequence.

D.2 Training with Uniform Representations

As an alternative to the multi-objective training paradigm, we explore a unified representation strategy in which both text and image modalities are represented as discrete tokens. Following the Chameleon framework Chameleon Team (2024), we treat the image data as sequences of discrete tokens obtained through a pre-trained VQ-VAE model Gafni et al. (2022). Specifically, each image is encoded into a fixed number of tokens (e.g., 1,024) by quantizing its latent features into a learned codebook. These tokens are then arranged sequentially, similar to the processing of text tokens, resulting in a uniform discrete representation across both modalities.

During training, both text and image tokens are processed using the same autoregressive objective, where the model learns to predict the next token in the sequence given all previous tokens. Formally, the training objective is:

$$\mathcal{L}_{\text{uniform}} = \mathbb{E}_{\mathbf{x}_{1:T}} \left[-\log P(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}) \right], \tag{6}$$

where $\mathbf{x}_{1:T}$ represents the interleaved sequence of text and image tokens. This objective allows the model to treat text and image data equivalently, unifying the training process across modalities while relying solely on an autoregressive loss. The use of discrete tokens for images simplifies the training procedure by removing the need for separate loss formulations, as in the diffusion-based approach. It also aligns with the inherent sequence-to-sequence nature of Mixture-of-Mamba, where the same modality-aware sparsity design can be applied seamlessly across the discrete text and image tokens.

Motivation and Robustness Testing. We include this alternative strategy to evaluate the robustness of our Mixture-of-Mamba architecture under different choices of training objectives and data representations. By experimenting with uniform discrete representations, we demonstrate that Mixture-of-Mamba consistently outperforms Mamba Dense models across various settings, including both continuous (multi-objective) and discrete (uniform) representations. This highlights the versatility of Mixture-of-Mamba and its ability to deliver performance gains regardless of the underlying choice of modality representations or training objectives.

E Ablation Study on Decoupling Components

To better understand the design choices underpinning Mixture-of-Mamba, we conduct an ablation study on the **Chameleon + Speech setting** at the **443M** scale. We evaluate the impact of decoupling four key components—$W_{\text{in-proj}}$ (❶), $W_{\text{x-proj}}$ (❷), $W_{\text{dt-proj}}$ (❸), and $W_{\text{out-proj}}$ (❹)—individually and in various combinations. This analysis enables us to test both individual and combined contributions to the model's overall performance.

The results show that decoupling components individually yields varying degrees of improvement, with performance gains ranging from **0.63%** ($W_{\text{out-proj}}$) to **1.22%** ($W_{\text{in-proj}}$). Interestingly, some components ($W_{\text{x-proj}}$ and $W_{\text{dt-proj}}$) exhibit minimal or even slightly negative impact when decoupled alone. However, decoupling multiple components in combination leads to significantly larger gains. For example, decoupling $W_{\text{in-proj}}$ and $W_{\text{out-proj}}$ (❶+❹) achieves a **2.20%** improvement, while decoupling three components (❶+❷+❹) further increases the gain to **3.11%**.

Most importantly, decoupling all four components simultaneously (❶+❷+❸+❹, Mixture-of-Mamba) achieves the largest improvement, with a performance gain of **3.80%** over the Mamba baseline. This result highlights a key observation: the gain from decoupling all components together exceeds the sum of individual gains, demonstrating a synergistic effect. The combination of all decoupled projections enables better parameter allocation across modalities, leading to more efficient and effective learning. In summary, the ablation study confirms that the design of Mixture-of-Mamba is both effective and interdependent. Decoupling all key components simultaneously is important to achieving the observed substantial performance gains.

Algorithm 2 Mixture-of-Mamba block

input $F_{in}, A, W_{in_proj}, W_{x_proj}, W_{dt_proj}, W_{out_proj}, b, M$
output F_{out}
$x, z \leftarrow \mathcal{M}(F_{in}, W_{in_proj}; M)$ ▷ Block starts
$u \leftarrow \text{SiLU}(\text{Conv1D}(x))$ ▷ [b,ℓ,d]
$\delta, B, C \leftarrow \mathcal{M}(u, W_{x_proj}; M)$ ▷ [b,ℓ,(r,n,n)]
$\Delta \leftarrow \log(1 + \exp((\mathcal{M}(\delta, W_{dt_proj}, b; M))))$
$\overline{A} \leftarrow \Delta * A$ ▷ [b,ℓ,d,n]
$\overline{B} \leftarrow \Delta * (u \times B)$ ▷ [b,ℓ,d,n]
$h = 0$ ▷ [b,d,n]
for $i = 0...N-1$ **do**
  $h = h * \overline{A_i} + \overline{B_i}$ ▷ [b,d,n]
  $y_i = h \cdot C_i$ ▷ [b,d]
end for
$o \leftarrow (y + u) * \text{SiLU}(z)$
$F_{out} \leftarrow \mathcal{M}(o, W_{out_proj}; M)$ ▷ Block ends

function $\mathcal{M}(X, W, b = \text{None}; M)$
  for each modality $m \in M$ **do**
    $I_m \leftarrow \{i : m_i = m\}$
    $X_m \leftarrow \{x_i : i \in I_m\}$
    $Y_m \leftarrow X_m W_m + b_m$
  end for
  return $Y \leftarrow \cup_{m \in M} Y_m$
end function

Ablation Study	Avg Training Loss (↓)	Performance Gain (%) (↑)
443M Mamba (without ❶❷❸❹)	3.3317	0% (baseline)
❶ (decouple W_{in_proj})	3.2916	1.22%
❷ (decouple W_{x_proj})	3.3580	-0.79%
❸ (decouple W_{dt_proj})	3.3525	-0.62%
❹ (decouple W_{out_proj})	3.3109	0.63%
❶+❷ (decouple W_{in_proj}, W_{x_proj})	3.2780	1.64%
❶+❸ (decouple W_{in_proj}, W_{dt_proj})	3.2687	1.93%
❶+❹ (decouple $W_{in_proj}, W_{out_proj}$)	3.2599	2.20%
❷+❸ (decouple W_{x_proj}, W_{dt_proj})	3.3214	0.31%
❷+❹ (decouple W_{x_proj}, W_{out_proj})	3.2829	1.49%
❸+❹ (decouple $W_{dt_proj}, W_{out_proj}$)	3.2509	2.48%
❶+❷+❸ (not decoupling W_{out_proj})	3.2593	2.22%
❶+❷+❹ (not decoupling W_{dt_proj})	3.2312	3.11%
❶+❸+❹ (not decoupling W_{x_proj})	3.2342	3.01%
❷+❸+❹ (not decoupling W_{in_proj})	3.2773	1.66%
❶+❷+❸+❹ (Mixture-of-Mamba)	**3.2096**	**3.80%**

Table 3: **Ablation study on the Chameleon + Speech setting.** This study evaluates the impact of decoupling individual components (**1**, **2**, **3**, **4**) and their combinations on model performance. The results demonstrate that decoupling all components (**1+2+3+4**, Mixture-of-Mamba) achieves the best performance with a **3.80%** gain over the Mamba baseline. Notably, the performance gain achieved by decoupling all components together exceeds the sum of gains from decoupling each component individually, highlighting the synergistic effect of combined decoupling. Green shading indicates positive performance gains, with the darkest green highlighting the best configuration.

Model Scale	Metric Category	Metric Name	Mamba Loss (↓)	Flex-Attention Transformer Loss (↓)	Mixture-of-Mamba Loss (↓)	Performance Gain over Mamba (%) (↑)	Relative Training FLOPs to Match Mamba (%) (↓)
163M	Image Metrics	Training Loss	0.2262	0.2250	*0.2199*	2.80%	49.21%
		CC12M Val. Loss	0.2295	0.2293	*0.2255*	1.74%	50.61%
	Text Metrics	Avg Training Loss	2.4702	*2.4424*	2.4690	0.05%	98.80%
		C4 Val. Loss	2.6917	*2.6862*	2.6912	0.02%	99.88%
		Wikipedia Val. Loss	2.1884	*2.1715*	2.1870	0.06%	99.81%
	Overall	Train Avg Loss	3.6014	*3.5674*	3.5685	0.91%	86.11%
760M	Image Metrics	Training Loss	0.2225	0.2213	*0.2172*	2.37%	37.76%
		CC12M Val. Loss	0.2272	0.2253	*0.2201*	3.13%	35.27%
	Text Metrics	Avg Training Loss	2.1394	*2.1253*	2.1353	0.19%	96.82%
		C4 Val. Loss	2.3593	*2.3559*	2.3555	0.16%	99.01%
		Wikipedia Val. Loss	1.8191	*1.8143*	1.8149	0.23%	99.11%
	Overall	Train Avg Loss	3.2519	3.2318	*3.2214*	0.94%	82.94%
1.4B	Image Metrics	Training Loss	0.2186	0.2221	*0.2138*	2.20%	34.76%
		CC12M Val. Loss	0.2264	0.2247	*0.2190*	3.29%	36.15%
	Text Metrics	Avg Training Loss	2.0761	*2.0673*	2.0737	0.12%	98.27%
		C4 Val. Loss	2.2726	2.2728	*2.2695*	0.13%	99.34%
		Wikipedia Val. Loss	1.7205	1.7218	*1.7164*	0.24%	99.30%
	Overall	Train Avg Loss	3.1693	3.1777	*3.1429*	0.84%	83.10%

Table 4: **Training and validation metrics across model scales in the Transfusion setting.** Loss values are reported for image and text modalities at three model sizes: **163M**, **760M**, and **1.4B**. Mixture-of-Mamba consistently achieves competitive or superior performance in image metrics and maintains strong text performance compared to Mamba Dense and Flex-Attention Transformer. The table also reports relative training FLOPs required for Mixture-of-Mamba and Flex-Attention Transformer to match Mamba's training dynamics, highlighting improved training efficiency. Best loss values in each row are highlighted.

Model Size	Hidden Dim.	Layers	Heads	Seq. Length	Batch Size/GPU	GPUs	Tokens/Batch	Steps
163M	768	16	12	4,096	4	56	1,048,576	250,000
760M	1,536	24	24	4,096	4	56	1,048,576	250,000
1.4B	2,048	24	16	4,096	2	128	1,048,576	250,000

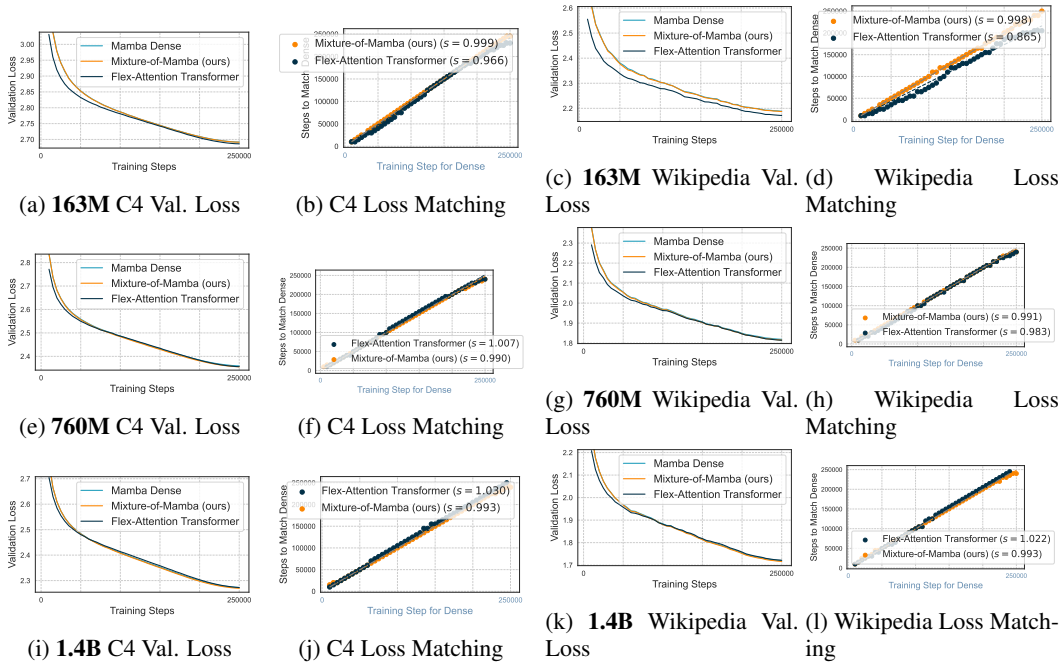
Table 5: **Architectural specifications and training configurations of models across different parameter scales (Transfusion setting).**

Model Size	Hidden Dim.	Layers	Heads	Seq. Length	Batch Size/GPU	GPUs	Tokens/Batch	Steps
37M	256	4	8	4,096	2	64	524,288	160,000
94M	512	8	8	4,096	2	64	524,288	160,000
443M	1,024	24	16	4,096	2	64	524,288	160,000
880M	1,536	24	24	4,096	2	64	524,288	120,000
1.5B	2,048	24	16	4,096	1	128	524,288	120,000

Table 6: **Architectural specifications and training configurations of models across different parameter scales (Chameleon setting and Chameleon+Speech setting).**

Model Scale	Metric Category	Metric Name	Mamba Loss (↓)	Mixture-of-Mamba Loss (↓)	Performance Gain (%) (↑)	Relative Training FLOPs to Match Mamba (%) (↓)
37M	Image Metrics	Training Loss	6.1308	5.9561	2.85%	25.90%
		Obelisc Val. Loss	5.2866	5.1124	3.29%	26.60%
		SSTK Val. Loss	6.6694	6.5023	2.51%	27.50%
	Text Metrics	Training Loss	3.6262	3.5175	3.00%	60.90%
		Obelisc Val. Loss	4.1244	4.0469	1.88%	64.80%
		SSTK Val. Loss	4.0417	3.9533	2.19%	57.50%
	Overall	Avg Training Loss	4.6607	4.5247	2.92%	50.70%
94M	Image Metrics	Training Loss	5.7609	5.6057	2.69%	35.70%
		Obelisc Val. Loss	4.9231	4.7683	3.14%	35.30%
		SSTK Val. Loss	6.3130	6.1652	2.34%	37.00%
	Text Metrics	Training Loss	3.0294	2.9414	2.90%	58.40%
		Obelisc Val. Loss	3.6016	3.5270	2.07%	62.60%
		SSTK Val. Loss	3.4109	3.2901	3.54%	61.40%
	Overall	Avg Training Loss	4.1577	4.0419	2.78%	49.80%
443M	Image Metrics	Training Loss	5.3558	5.1703	3.46%	33.40%
		Obelisc Val. Loss	4.5258	4.3546	3.78%	35.10%
		SSTK Val. Loss	5.9179	5.7471	2.89%	35.30%
	Text Metrics	Training Loss	2.4637	2.3864	3.14%	62.00%
		Obelisc Val. Loss	3.0544	2.9820	2.37%	66.70%
		SSTK Val. Loss	2.7569	2.6250	4.78%	54.70%
	Overall	Avg Training Loss	3.6584	3.5364	3.33%	47.90%
880M	Image Metrics	Training Loss	5.2260	5.1201	2.03%	48.40%
		Obelisc Val. Loss	4.4127	4.3105	2.32%	49.30%
		SSTK Val. Loss	5.7987	5.6986	1.73%	50.50%
	Text Metrics	Training Loss	2.3073	2.2438	2.75%	65.60%
		Obelisc Val. Loss	2.8886	2.8313	1.99%	72.80%
		SSTK Val. Loss	2.5483	2.4548	3.67%	67.90%
	Overall	Avg Training Loss	3.5130	3.4320	2.31%	58.30%
1.5B	Image Metrics	Training Loss	5.1892	5.0591	2.51%	42.50%
		Obelisc Val. Loss	4.3692	4.2510	2.71%	44.50%
		SSTK Val. Loss	5.7546	5.6335	2.10%	44.60%
	Text Metrics	Training Loss	2.2284	2.1614	3.01%	65.40%
		Obelisc Val. Loss	2.8020	2.7393	2.24%	71.60%
		SSTK Val. Loss	2.4614	2.3455	4.71%	62.10%
	Overall	Avg Training Loss	3.4602	3.3670	2.69%	54.70%

Table 7: **Training and validation metrics across model scales in the Chameleon setting.** In this setting, both image and text modalities are represented as discrete tokens. Mixture-of-Mamba achieves substantial performance improvements over Mamba Dense, with the **image modality** showing the largest gains across all five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. Notably, the **text modality** also exhibits significant improvements, in contrast to the Transfusion setting where text gains were more modest. These results further highlight the effectiveness and efficiency of Mixture-of-Mamba, which consistently achieves strong performance with reduced relative training FLOPs.

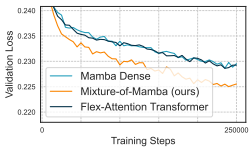


(a) **163M** C4 Val. Loss (b) C4 Loss Matching (c) **163M** Wikipedia Val. Loss (d) Wikipedia Loss Matching

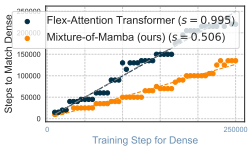
(e) **760M** C4 Val. Loss (f) C4 Loss Matching (g) **760M** Wikipedia Val. Loss (h) Wikipedia Loss Matching

(i) **1.4B** C4 Val. Loss (j) C4 Loss Matching (k) **1.4B** Wikipedia Val. Loss (l) Wikipedia Loss Matching

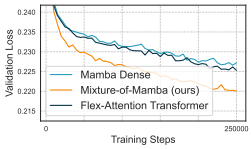
Figure 3: Validation loss and loss matching for text modality across model scales (**C4** and **Wikipedia** datasets) during multi-modal pretraining in the **Transfusion setting**. Results are shown for Mixture-of-Mamba, Mamba Dense, and Flex-Attention Transformer at three model scales: **163M**, **760M**, and **1.4B**. (**a, e, i**) Validation loss on the **C4 dataset** shows that Mixture-of-Mamba achieves comparable performance at **163M** and performs marginally better than Mamba Dense and Flex-Attention Transformer at the **760M** and **1.4B** scales. (**b, f, j**) Loss matching for C4 demonstrates that Mixture-of-Mamba reaches similar or slightly lower loss values at earlier training steps compared to Mamba Dense. (**c, g, k**) Validation loss on the **Wikipedia dataset** follows a similar trend, with Mixture-of-Mamba showing marginal improvements at the **760M** and **1.4B** scales. (**d, h, l**) Loss matching for Wikipedia illustrates efficient training dynamics, with Mixture-of-Mamba aligning closely with Flex-Attention Transformer while reaching comparable or slightly lower loss values than Mamba Dense. Overall, Mixture-of-Mamba demonstrates moderate improvements over both baselines at the larger scales (**760M** and **1.4B**).



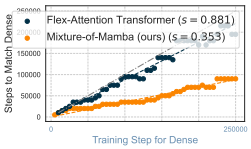
(a) **163M** Image Val. Loss



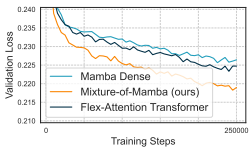
(b) Image Val. Loss Matching



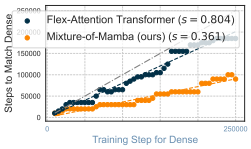
(c) **760M** Image Val. Loss



(d) Image Val. Loss Matching

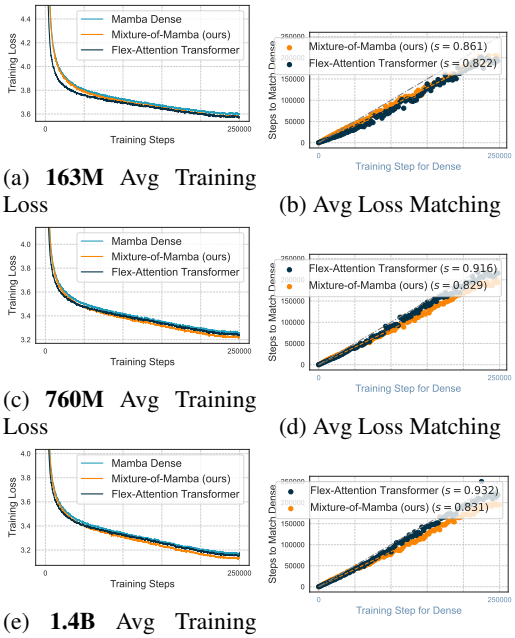


(e) **1.4B** Image Val. Loss



(f) Image Val. Loss Matching

Figure 4: Image validation loss and loss matching on the **CC12M dataset** across three model scales: **163M**, **760M**, and **1.4B** during multi-modal pretraining in the **Transfusion setting**. (**a, c, e**) Validation loss curves show that Mixture-of-Mamba achieves substantially lower image validation loss compared to Mamba Dense and Flex-Attention Transformer across all scales, with the improvement becoming more pronounced as model size increases. (**b, d, f**) Loss matching curves demonstrate that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, highlighting improved training efficiency. Overall, Mixture-of-Mamba achieves large improvements in image validation loss on the **CC12M dataset**, showcasing its effectiveness in the image modality.



(a) **163M** Avg Training Loss

(b) Avg Loss Matching

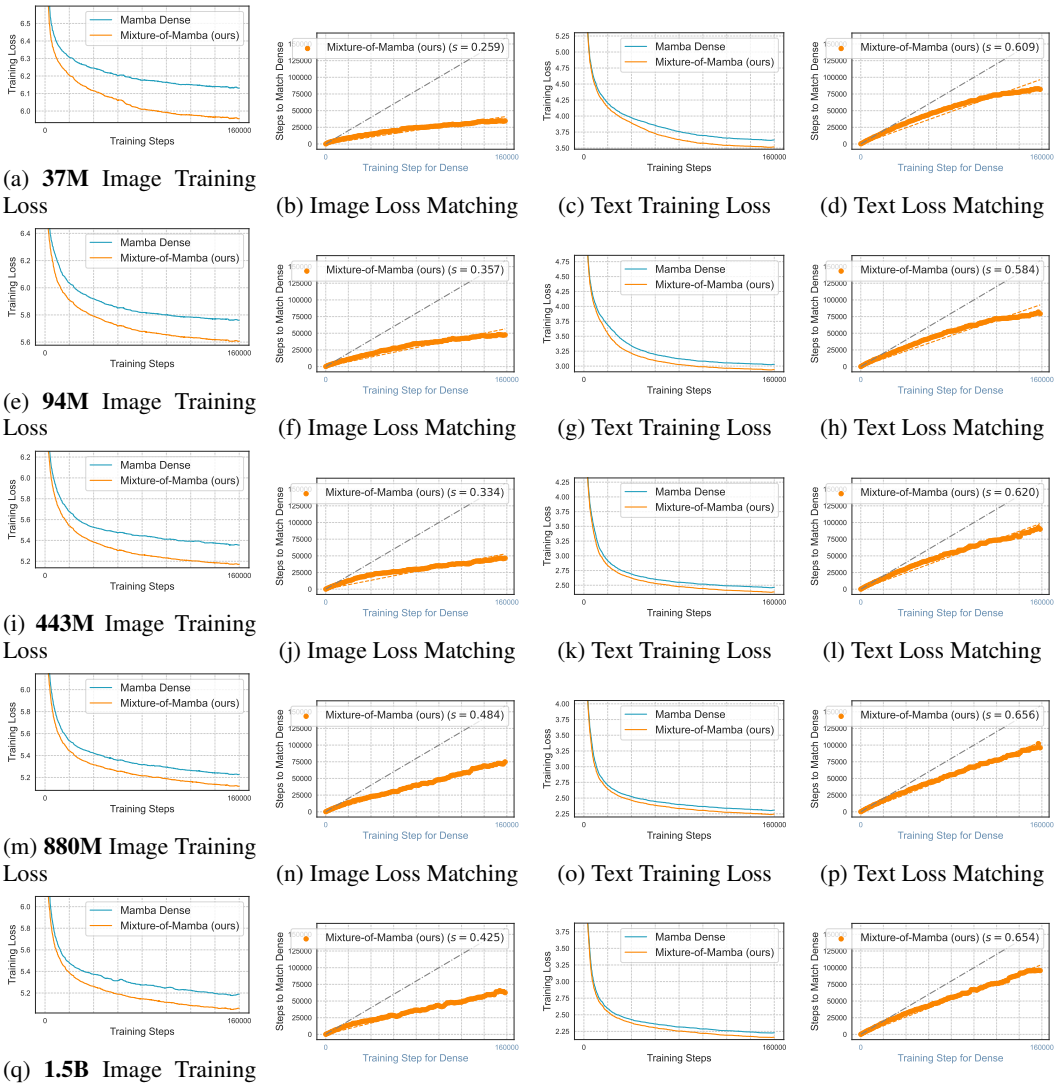
(c) **760M** Avg Training Loss

(d) Avg Loss Matching

(e) **1.4B** Avg Training Loss

(f) Avg Loss Matching

Figure 5: Overall training loss and loss matching during multi-modal pretraining in the **Transfusion setting**. Results are shown for Mixture-of-Mamba, Mamba Dense, and Flex-Attention Transformer at three model scales: **163M**, **760M**, and **1.4B**. (**a, c, e**) Training loss averaged across the image and text modalities demonstrates that Mixture-of-Mamba achieves substantial improvements over Mamba Dense, with a notable reduction in training loss across all scales. (**b, d, f**) Loss matching results show that Mixture-of-Mamba and Flex-Attention Transformer reach the same loss values at earlier training steps compared to Mamba Dense, highlighting improved training efficiency. *Note:* The image loss in the Transfusion setting corresponds to the diffusion loss, which is of smaller magnitude compared to the cross-entropy loss in the text modality. Overall, Mixture-of-Mamba demonstrates significant gains in training loss and efficiency across multi-modal pretraining.



(a) **37M** Image Training Loss (b) Image Loss Matching (c) Text Training Loss (d) Text Loss Matching

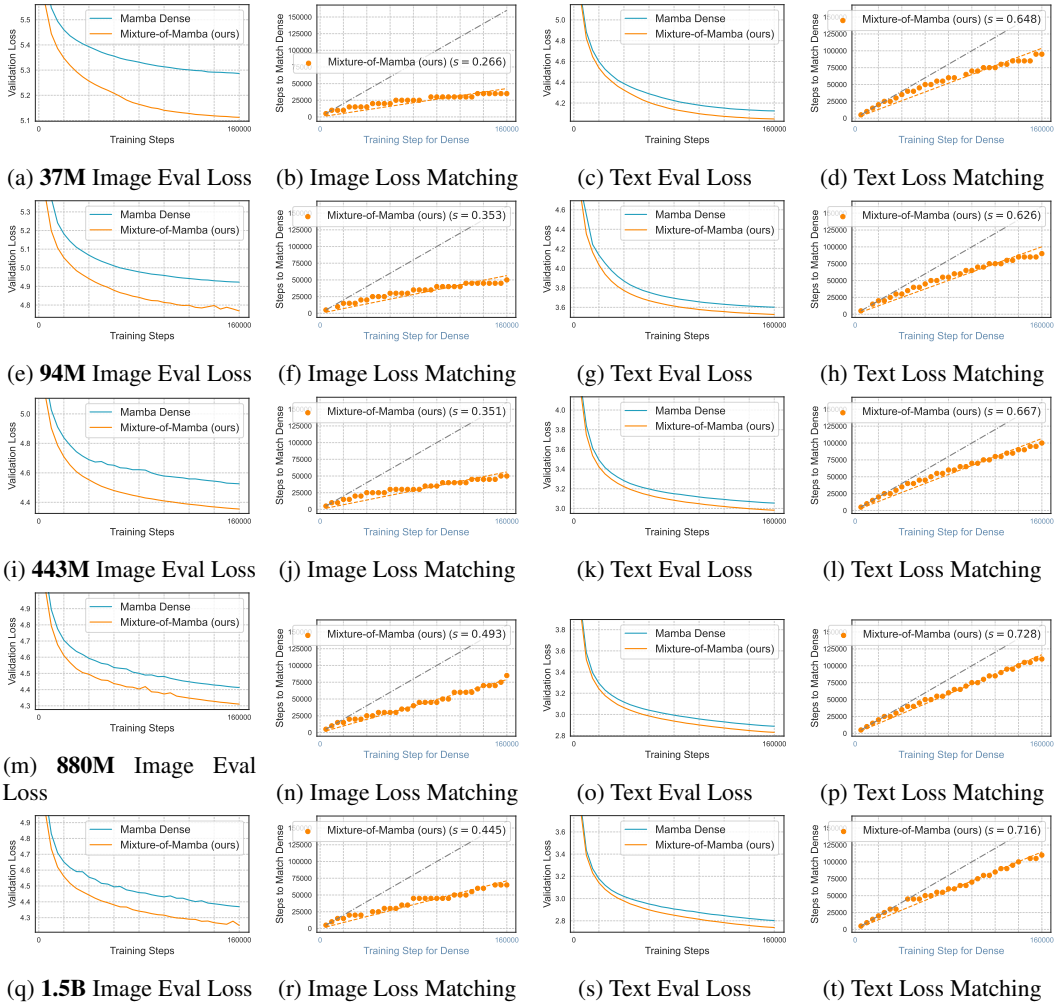
(e) **94M** Image Training Loss (f) Image Loss Matching (g) Text Training Loss (h) Text Loss Matching

(i) **443M** Image Training Loss (j) Image Loss Matching (k) Text Training Loss (l) Text Loss Matching

(m) **880M** Image Training Loss (n) Image Loss Matching (o) Text Training Loss (p) Text Loss Matching

(q) **1.5B** Image Training Loss (r) Image Loss Matching (s) Text Training Loss (t) Text Loss Matching

Figure 6: **Modality-specific pre-training loss and step matching plots across model scales (Chameleon setting).** Training loss and loss matching are reported for image and text modalities across five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. **(a, e, i, m, q)** Image training loss shows significant improvements for Mixture-of-Mamba (orange), which consistently achieves lower loss compared to Mamba Dense (cyan) across all scales. **(b, f, j, n, r)** Image loss matching compares the training dynamics and shows that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, highlighting its improved efficiency. **(c, g, k, o, s)** Text training loss demonstrates competitive performance, with Mixture-of-Mamba achieving slightly lower loss values compared to Mamba Dense. **(d, h, l, p, t)** Text loss matching illustrates that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, reflecting its efficient training dynamics. Overall, in the **Chameleon setting**, Mixture-of-Mamba achieves consistent improvements in the image modality, with substantial computational savings, while also demonstrating meaningful gains in the text modality.



(a) **37M** Image Eval Loss (b) Image Loss Matching (c) Text Eval Loss (d) Text Loss Matching

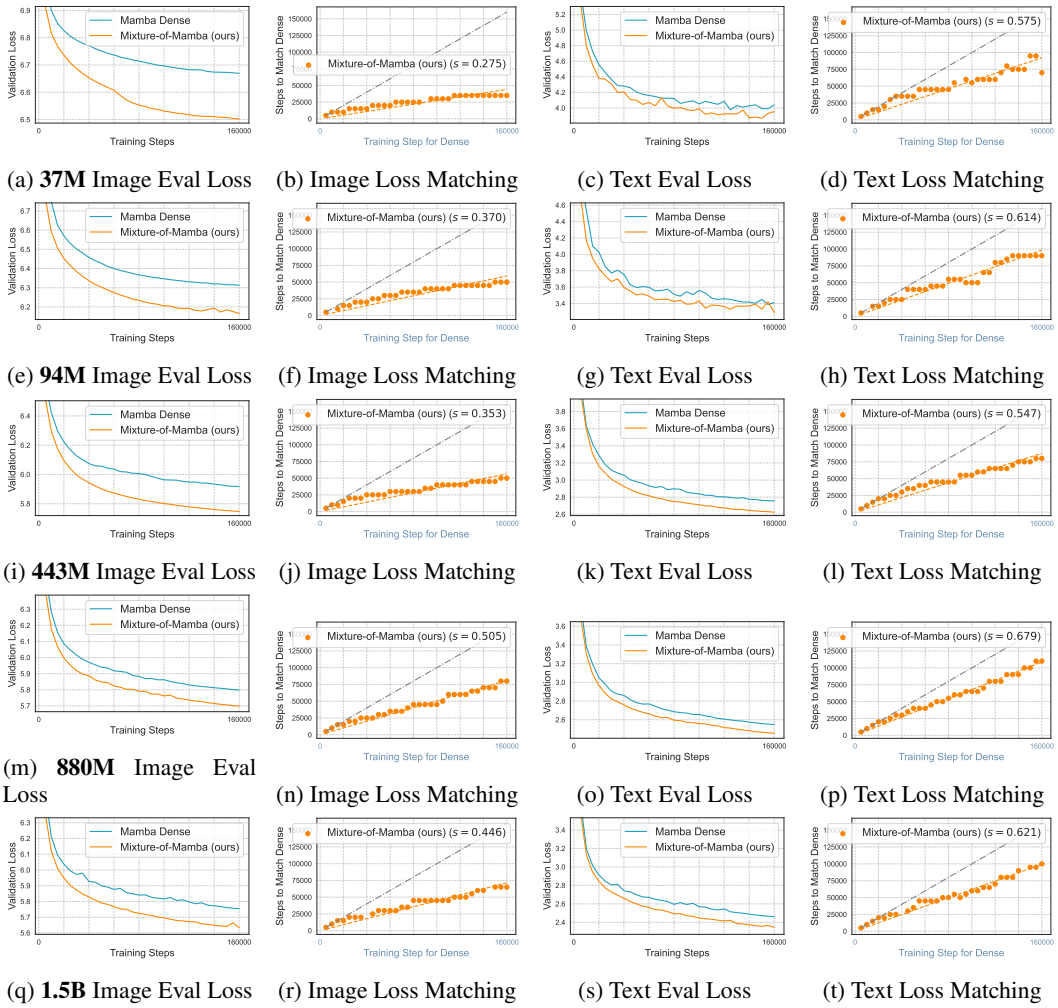
(e) **94M** Image Eval Loss (f) Image Loss Matching (g) Text Eval Loss (h) Text Loss Matching

(i) **443M** Image Eval Loss (j) Image Loss Matching (k) Text Eval Loss (l) Text Loss Matching

(m) **880M** Image Eval Loss (n) Image Loss Matching (o) Text Eval Loss (p) Text Loss Matching

(q) **1.5B** Image Eval Loss (r) Image Loss Matching (s) Text Eval Loss (t) Text Loss Matching

Figure 7: **Training and evaluation losses for image and text modalities across model scales in the Chameleon setting on the Obelisc dataset.** Results are shown for Mixture-of-Mamba and Mamba Dense across five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. **(a, e, i, m, q)** Image evaluation loss demonstrates consistent improvements for Mixture-of-Mamba (orange), achieving lower loss compared to Mamba Dense (cyan) across all scales. **(b, f, j, n, r)** Image loss matching shows that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, reflecting its improved training efficiency. **(c, g, k, o, s)** Text evaluation loss indicates competitive results for Mixture-of-Mamba, achieving lower losses relative to Mamba Dense. **(d, h, l, p, t)** Text loss matching highlights that Mixture-of-Mamba reaches the same loss values at earlier training steps, further demonstrating its efficiency in the text modality. Overall, Mixture-of-Mamba achieves strong and consistent improvements in both image and text modalities across all model scales in the Chameleon setting evaluated on the Obelisc dataset.



(a) **37M** Image Eval Loss (b) Image Loss Matching (c) Text Eval Loss (d) Text Loss Matching

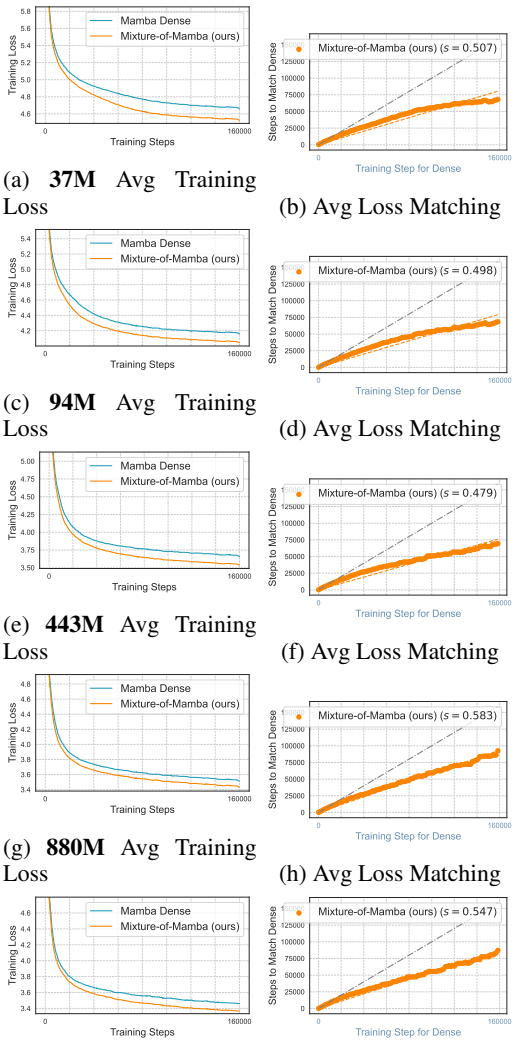
(e) **94M** Image Eval Loss (f) Image Loss Matching (g) Text Eval Loss (h) Text Loss Matching

(i) **443M** Image Eval Loss (j) Image Loss Matching (k) Text Eval Loss (l) Text Loss Matching

(m) **880M** Image Eval Loss (n) Image Loss Matching (o) Text Eval Loss (p) Text Loss Matching

(q) **1.5B** Image Eval Loss (r) Image Loss Matching (s) Text Eval Loss (t) Text Loss Matching

Figure 8: **Training and evaluation losses for image and text modalities across model scales in the Chameleon setting on the Shutterstock dataset.** Results are shown for Mixture-of-Mamba and Mamba Dense across five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. **(a, e, i, m, q)** Image evaluation loss demonstrates consistent improvements for Mixture-of-Mamba (orange), achieving lower loss compared to Mamba Dense (cyan) across all scales. **(b, f, j, n, r)** Image loss matching shows that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, reflecting its improved training efficiency. **(c, g, k, o, s)** Text evaluation loss indicates competitive results for Mixture-of-Mamba, achieving lower losses relative to Mamba Dense. **(d, h, l, p, t)** Text loss matching highlights that Mixture-of-Mamba reaches the same loss values at earlier training steps, further demonstrating its efficiency in the text modality. Overall, Mixture-of-Mamba achieves strong and consistent improvements in both image and text modalities across all model scales in the Chameleon setting evaluated on the Shutterstock dataset.



(a) **37M** Avg Training Loss (b) Avg Loss Matching

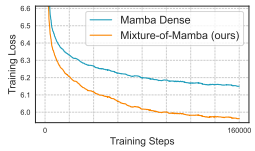
(c) **94M** Avg Training Loss (d) Avg Loss Matching

(e) **443M** Avg Training Loss (f) Avg Loss Matching

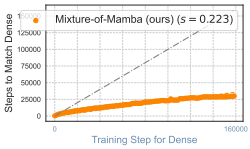
(g) **880M** Avg Training Loss (h) Avg Loss Matching

(i) **1.5B** Avg Training Loss (j) Avg Loss Matching

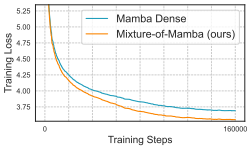
Figure 9: **Average training loss and step matching plots across model scales in the Chameleon setting.** Results are shown for Mixture-of-Mamba and Mamba Dense across five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. **(a, c, e, g, i)** Average training loss (across image and text modalities) demonstrates consistent reductions for Mixture-of-Mamba (orange), achieving lower loss values compared to Mamba Dense (cyan) at all model scales. **(b, d, f, h, j)** Average loss matching plots highlight that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, reflecting improved training efficiency. Overall, Mixture-of-Mamba consistently reduces average training loss and achieves more efficient convergence across all model scales in the Chameleon setting.



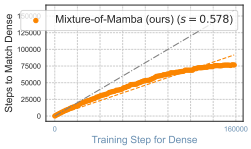
(a) **37M** Image Training Loss



(b) Image Loss Matching

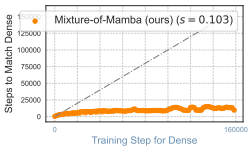


(c) Text Training Loss

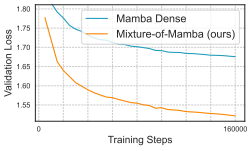


(d) Text Loss Matching

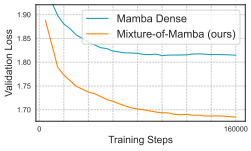
(e) **37M** Speech Training Loss



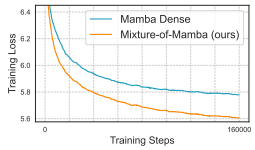
(f) Speech Loss Matching



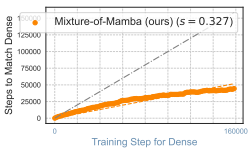
(g) Speech Eval Loss: LL60K



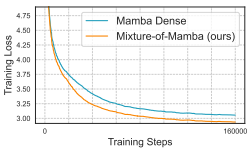
(h) Speech Eval Loss: PPL30K



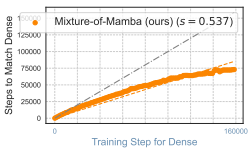
(i) **94M** Image Training Loss



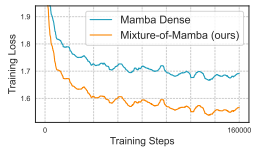
(j) Image Loss Matching



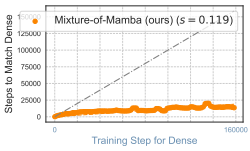
(k) Text Training Loss



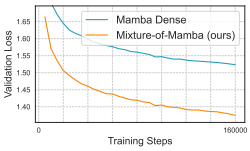
(l) Text Loss Matching



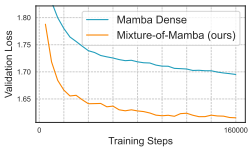
(m) **94M** Speech Training Loss



(n) Speech Loss Matching

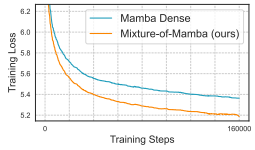


(o) Speech Eval Loss: LL60K

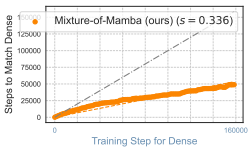


(p) Speech Eval Loss: PPL30K

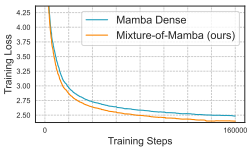
Figure 10: **Training and evaluation losses for image, text, and speech modalities (37M and 94M scales) in the Chameleon+Speech setting.** Results are reported for Mixture-of-Mamba and Mamba Dense. **(a, e, i)** Image training loss demonstrates that Mixture-of-Mamba (orange) achieves consistently lower loss compared to Mamba Dense (cyan). **(b, f, j)** Image loss matching highlights Mixture-of-Mamba's ability to reach the same loss values at earlier training steps, showing improved training efficiency. **(c, g, k)** Text training loss shows competitive results for Mixture-of-Mamba, improving over Mamba Dense. **(d, h, l)** Text loss matching confirms Mixture-of-Mamba's ability to reach the same loss values at earlier training steps, showing improved training efficiency. **(e, m)** Speech training loss highlights significant improvements in speech modality performance. **(f, n)** Speech loss matching shows efficient learning dynamics for Mixture-of-Mamba. **(g, o)** Speech evaluation loss on LL60K confirms notable performance gains, and **(h, p)** Speech evaluation loss on PPL30K further highlights the efficiency of Mixture-of-Mamba.



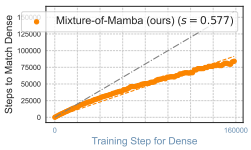
(a) **443M** Image Training Loss



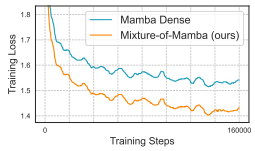
(b) Image Loss Matching



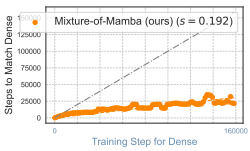
(c) Text Training Loss



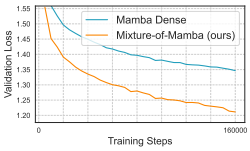
(d) Text Loss Matching



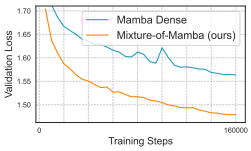
(e) **443M** Speech Training Loss



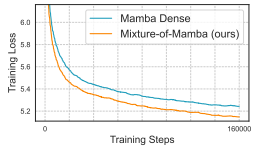
(f) Speech Loss Matching



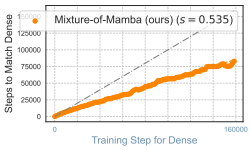
(g) Speech Eval Loss: LL60K



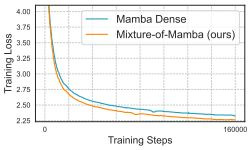
(h) Speech Eval Loss: PPL30K



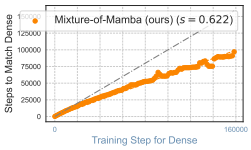
(i) **880M** Image Training Loss



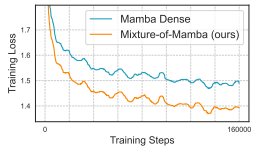
(j) Image Loss Matching



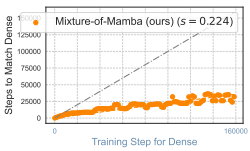
(k) Text Training Loss



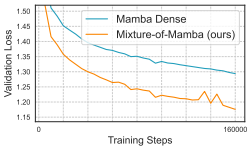
(l) Text Loss Matching



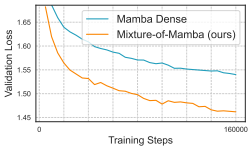
(m) **880M** Speech Training Loss



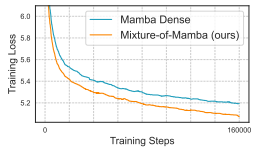
(n) Speech Loss Matching



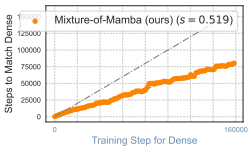
(o) Speech Eval Loss: LL60K



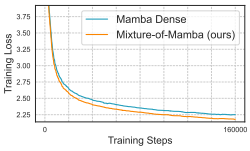
(p) Speech Eval Loss: PPL30K



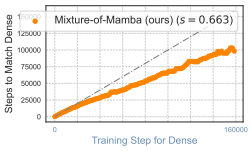
(q) **1.5B** Image Training Loss



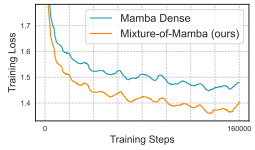
(r) Image Loss Matching



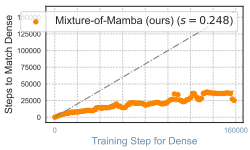
(s) Text Training Loss



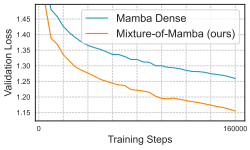
(t) Text Loss Matching



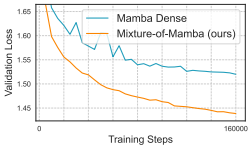
(u) **1.5B** Speech Training Loss



(v) Speech Loss Matching

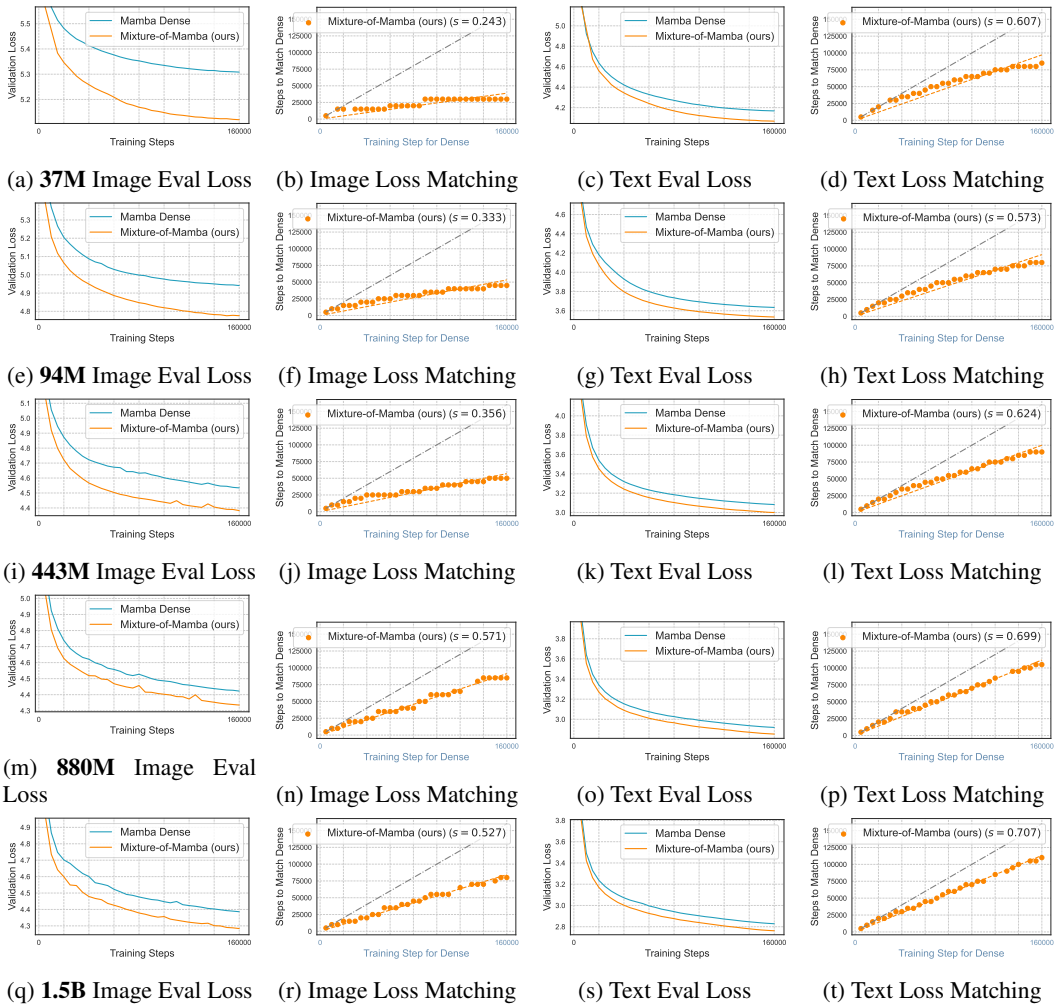


(w) Speech Eval Loss: LL60K



(x) Speech Eval Loss: PPL30K

Figure 11: **Training and evaluation losses for image, text, and speech modalities (443M, 880M, and 1.5B scales) in the Chameleon+Speech setting.** Results are reported for Mixture-of-Mamba and Mamba Dense. **(a, i, q)** Image training loss demonstrates that Mixture-of-Mamba (orange) consistently outperforms Mamba Dense (cyan) across larger scales. **(b, j, r)** Image loss matching highlights improved training efficiency for Mixture-of-Mamba, reaching the same loss values at earlier training steps. **(c, k, s)** Text training loss shows Mixture-of-Mamba achieving better performance. **(d, l, t)** Text loss matching further demonstrates efficient learning dynamics. **(e, m, u)** Speech training loss confirms substantial gains for Mixture-of-Mamba in the speech modality, consistent across model scales. **(f, n, v)** Speech loss matching illustrates the improved efficiency of Mixture-of-Mamba across scales. **(g, o, w)** Speech evaluation loss on LL60K highlights consistent improvements, while **(h, p, x)** Speech evaluation loss on PPL30K demonstrates notable gains and efficient performance across scales.



(a) **37M** Image Eval Loss (b) Image Loss Matching (c) Text Eval Loss (d) Text Loss Matching

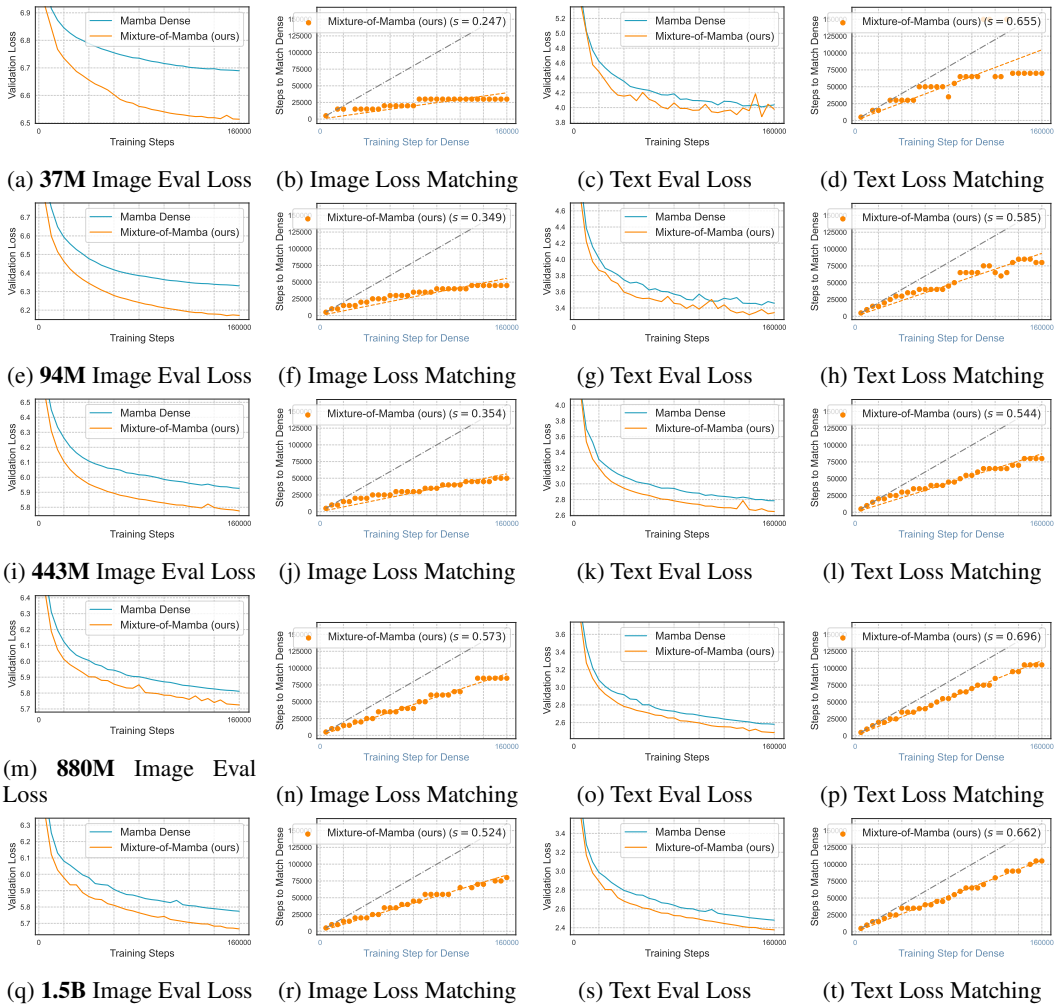
(e) **94M** Image Eval Loss (f) Image Loss Matching (g) Text Eval Loss (h) Text Loss Matching

(i) **443M** Image Eval Loss (j) Image Loss Matching (k) Text Eval Loss (l) Text Loss Matching

(m) **880M** Image Eval Loss (n) Image Loss Matching (o) Text Eval Loss (p) Text Loss Matching

(q) **1.5B** Image Eval Loss (r) Image Loss Matching (s) Text Eval Loss (t) Text Loss Matching

Figure 12: **Training and validation losses for image and text modalities across model scales in the Chameleon+Speech setting evaluated on the Obelisc dataset.** Results are shown for Mixture-of-Mamba and Mamba Dense across five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. **(a, e, i, m, q)** Image evaluation loss demonstrates consistent gains for Mixture-of-Mamba (orange) over Mamba Dense (cyan), even with the inclusion of the speech modality. **(b, f, j, n, r)** Image loss matching shows that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, highlighting improved efficiency. **(c, g, k, o, s)** Text evaluation loss indicates consistent reductions for Mixture-of-Mamba relative to Mamba Dense across all scales. **(d, h, l, p, t)** Text loss matching illustrates that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, maintaining its efficiency in the text modality. Overall, Mixture-of-Mamba achieves consistent improvements in both image and text modalities while maintaining its efficiency, even with the addition of the **speech modality**. These results confirm the robustness of Mixture-of-Mamba in multi-modal settings.



(a) **37M** Image Eval Loss (b) Image Loss Matching (c) Text Eval Loss (d) Text Loss Matching

(e) **94M** Image Eval Loss (f) Image Loss Matching (g) Text Eval Loss (h) Text Loss Matching

(i) **443M** Image Eval Loss (j) Image Loss Matching (k) Text Eval Loss (l) Text Loss Matching

(m) **880M** Image Eval Loss (n) Image Loss Matching (o) Text Eval Loss (p) Text Loss Matching

(q) **1.5B** Image Eval Loss (r) Image Loss Matching (s) Text Eval Loss (t) Text Loss Matching

Figure 13: **Training and validation losses for image and text modalities across model scales in the Chameleon+Speech setting evaluated on the Shutterstock dataset.** Results are shown for Mixture-of-Mamba and Mamba Dense across five model scales: **37M**, **94M**, **443M**, **880M**, and **1.5B**. **(a, e, i, m, q)** Image evaluation loss demonstrates consistent gains for Mixture-of-Mamba (orange) over Mamba Dense (cyan), even with the inclusion of the speech modality. **(b, f, j, n, r)** Image loss matching shows that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, highlighting improved efficiency. **(c, g, k, o, s)** Text evaluation loss indicates consistent reductions for Mixture-of-Mamba relative to Mamba Dense across all scales. **(d, h, l, p, t)** Text loss matching illustrates that Mixture-of-Mamba reaches the same loss values at earlier training steps compared to Mamba Dense, maintaining its efficiency in the text modality. Overall, Mixture-of-Mamba achieves consistent improvements in both image and text modalities while maintaining its efficiency, even with the addition of the **speech modality**. These results confirm the robustness of Mixture-of-Mamba in multi-modal settings.