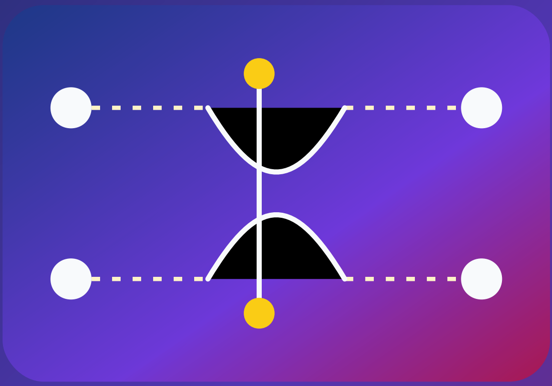

<h3>Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models · Concept Sketch</h3>

<p>Late-fusion inspired view showing coordinated yet modality-specific streams.</p>

Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models

Authors: Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, Xi Victoria Lin

Year: 2025
Venue: Transactions on Machine Learning Research (TMLR)

1. Classification

- **Domain Category:**
  - Vision / VLM / Multimodal FM
  - The paper proposes a new architecture for unified multi-modal generation and understanding (text, images, and speech) in large foundation models.
- **FM Usage Type:**
  - Core FM development + Multimodal FM or cross-modal integration
- **Key Modalities:**
  - Text (language tokens).
  - Images (discrete image tokens for generation and understanding).
  - Speech (discrete speech tokens in some settings).

2. Executive Summary

Mixture-of-Transformers (MoT) is a sparse multi-modal transformer architecture designed to make large, unified foundation models for text, images, and speech much more computationally efficient. Instead of running a single dense transformer over all modalities, MoT keeps global self-attention over the full mixed sequence but **decouples all non-embedding parameters by modality**: feed-forward networks, attention projections, and layer norms are specialized for each modality while sharing the same FLOP budget as the dense baseline. The authors evaluate MoT in several settings, including Chameleon-style autoregressive text–image and text–image–speech generation and Transfusion-style models that combine autoregressive text generation with diffusion-based image generation. Across scales from tens of millions to billions of parameters, MoT consistently matches or exceeds dense baselines while using **40–60% of the pretraining FLOPs**, and delivers substantial wall-clock speedups. The paper also studies modality separation, leave-one-modality-out experiments, and hybrid models that mix MoT with MoE-style experts, showing that modality-aware sparsity is a stable and effective alternative to learned routing. For a new grad student, MoT is a valuable example of how to introduce structured sparsity into multimodal transformers without sacrificing architectural simplicity.

3. Problem Setup and Motivation

- **Scientific / practical problem:**
    - Build **unified multi-modal foundation models** that can jointly process and generate text, images, and

    speech, but do so with **manageable compute budgets**.
  - Reduce the training and inference costs of early-fusion multimodal transformers (like Chameleon-style models) without giving up performance.
- **Why this is hard:**
  - **Dense transformers scale poorly** when extended from text-only LLMs to multi-modal settings; adding image and speech tokens massively increases sequence lengths and token diversity.
  - Different modalities have **conflicting optimization dynamics** and live in very different regions of representation space, so a single set of shared parameters may be suboptimal.
  - Mixture-of-Experts (MoE) architectures introduce routing instability, load-balancing overhead, and complex bi-level optimization; they are powerful but hard to train robustly at scale.
  - Any new architecture must preserve **implementation simplicity** and FLOP accounting so that practitioners can reason about cost vs. quality.

4. Data and Modalities

MoT is evaluated in multiple multi-modal scenarios built on existing benchmarks and systems:

- **Chameleon setting (text–image, text–image–speech):**
  - Autoregressive generation over interleaved text and image tokens; extended to include discrete speech tokens in a three-modality setup.

  - Trained on large-scale text–image and text–image–speech datasets similar to those used for prior Chameleon models.
- **Transfusion setting (text + diffusion images):**
  - Text is modeled autoregressively as in standard LLMs.
  - Images are modeled with diffusion-based objectives, using latent image representations.
- **Modalities:**
  - Text: natural language prompts, captions, and conversational context.
  - Images: discrete or latent tokens used for generation and captioning benchmarks.
  - Speech: discrete speech tokens for audio generation and understanding (in the Chameleon+Speech experiments).
- **Preprocessing / representation:**
  - All modalities are converted into a **single mixed token sequence**, with tokens tagged by modality so that MoT can apply modality-specific parameters while keeping shared self-attention over the whole sequence.
  - For diffusion images, latent representations and timesteps are embedded following standard diffusion-transformer practices.

5. Model / Foundation Model

- **Model Type:**
  - Sparse transformer architecture (**Mixture-of-Transformers**) with modality-specific parameters but dense-style global self-attention.
- **Is it a new FM or an existing one?**
  - MoT is a **new architectural pattern** that can be plugged into existing multimodal settings (e.g., Chameleon, Transfusion) while preserving their training objectives.
- **Key components and innovations:**

Aspect	Details
Sparsity mechanism	Modality-aware sparsity over all non-embedding parameters (FFNs, attention matrices, layer norms).
Shared attention	Full self-attention over the mixed multi-modal sequence; no routing-based sparsity in attention.

Aspect	Details
Parameter decoupling	For each modality, separate parameter sets for projections and FFNs, selected by token modality.
Compatibility	Drop-in replacement for dense transformers in Chameleon and Transfusion architectures.
Hybrid models	Combining MoT with MoE-4x to explore complementary benefits of expert routing and modality sparsity.

- **Training setup (high level):**
  - Pretrain **13 models** of various sizes (37M–7B, plus hybrid architectures) across multiple Chameleon and Transfusion settings.

    - Objectives:
        - Autoregressive next-token prediction for text and image tokens (Chameleon).
        - Autoregressive text + diffusion-based image objectives (Transfusion).
    - Compute measured in FLOPs and wall-clock time; experiments run on multi-GPU clusters (e.g., AWS p4de instances with A100s).

6. Multimodal / Integration Aspects (If Applicable)

MoT directly targets multi-modal integration in unified transformers:

- **Modalities integrated:**
    - Text, images, and (in some experiments) speech, all represented as tokens in one interleaved sequence.
- **How they are integrated:**
    - The model applies **global self-attention** across all tokens regardless of modality, enabling cross-modal interactions at every layer.
    - For each token, a simple rule based on its modality selects which FFN, attention projections, and layer norms to use; this is **rule-based routing by modality**, not learned MoE routing.
    - This yields a sparse model where only a subset of parameters are active for each token, but the computational graph (FLOPs) matches a dense transformer.

- **Why this integration is useful / new capabilities:**
    - Maintains the strengths of early-fusion multimodal transformers (rich cross-modal attention) while **reducing compute** and avoiding MoE training instability.
    - Makes it feasible to train unified multi-modal foundation models at larger scales and on more complex objectives (e.g., mixed autoregressive + diffusion) with limited resources.
    - Provides a clean baseline architecture for future multi-modal FMs that want structured sparsity without heavy routing machinery.

7. Experiments and Results

- **Settings:**
    - **Chameleon (text–image, text–image–speech)** for unified autoregressive generation.
    - **Transfusion (text + diffusion images)** for mixed-objective training.
    - Additional analyses of modality separation, leave-one-modality-out behavior, and systems aspects (throughput, scaling).
- **Baselines:**
    - Dense Chameleon and Transfusion transformers at comparable parameter scales.
    - Mixture-of-Experts (MoE-4x) variants that increase parameter count via expert routing.
- **Key findings (trends):**
    - In Chameleon 7B, MoT **matches dense performance** on text and image metrics while using

    only **55.8% of pretraining FLOPs**.
  - When extended to text–image–speech, MoT reaches **speech performance comparable to the dense baseline** with ~37% of the FLOPs for that modality.
  - In the Transfusion setting, a **760M MoT** outperforms a **1.4B dense baseline** on image generation and captioning metrics while using half the training and inference FLOPs; a 7B MoT matches the 7B dense model with roughly one third of the FLOPs for the image modality.
  - System profiling shows that MoT achieves the same image quality in **47.2% of the wall-clock time** and similar text quality in **75.6% of the time** on A100 clusters.
  - Compared to MoE-4x, MoT often provides better or more stable performance at similar or lower FLOP budgets.

8. Strengths, Limitations, and Open Questions

Strengths:

- Simple, **modality-aware sparsity mechanism** that integrates cleanly into standard transformers.
- Demonstrated **compute savings** (FLOPs and wall-clock) without sacrificing performance across several challenging multi-modal setups.
- Extensive empirical evaluation across model sizes and tasks, including systems-level profiling.

- Provides a practical alternative to MoE that avoids routing instability and complex load balancing.

Limitations:

- Still focused on a relatively small set of modalities (text, images, speech); does not cover more exotic or structured modalities (e.g., audio waveforms, tabular EHR, 3D point clouds).
- Requires modality labels for tokens; does not explore more fine-grained routing within a modality (e.g., by region or task).
- Results are tied to specific training infrastructures and datasets; real-world deployment costs may differ.
- The paper does not deeply explore how MoT interacts with instruction tuning, long-context training, or reinforcement learning, which are important in practice.

Open questions and future directions:

1. How well does MoT extend to **more modalities and tasks**, such as video, 3D data, or structured signals like EHR?
2. Can we combine **modality-aware sparsity with learned experts** in a principled way, getting the best of MoT and MoE while keeping training stable?
3. How does MoT behave under heavy **instruction tuning or RLHF**, where gradients can be noisy and task distributions shift?
4. Could similar modality-specific parameter decoupling be applied to **encoder–decoder or diffusion-only architectures** in a clean way?
5. What are the implications of MoT-style specialization for **interpretability**, e.g., understanding what each modality-specific block has learned?

9. Context and Broader Impact

- **Position in the landscape:**
    - MoT sits in the line of work on **unified multi-modal transformers** (e.g., Chameleon, Transfusion, integrated transformer-diffusion models) and offers a new way to scale them efficiently.
    - It complements MoE-style sparse methods by providing a simpler, modality-aware alternative that still preserves early-fusion attention.
- **Relation to well-known ideas:**
    - Borrowing the idea of **sparsity per token** from MoE, but replacing learned routing with deterministic routing by modality.
    - Conceptually similar to having **per-modality adapters everywhere** in the transformer, but integrated as full parameter sets rather than small adapter layers.
- **Why it is a useful reference:**
    - For researchers designing multi-modal FMs, MoT shows how to trade off computation and flexibility without sacrificing architectural clarity.
    - For systems and efficiency work, it offers a realistic case study in measuring FLOPs and wall-clock savings in large multimodal training runs.

10. Key Takeaways (Bullet Summary)

- **Problem:**
  - Early-fusion multimodal transformers for unified text–image–speech generation are extremely compute-hungry; MoE-style sparsity is powerful but unstable and complex.
- **Method / model:**
  - Mixture-of-Transformers (MoT) introduces **modality-aware sparsity** by decoupling non-embedding transformer parameters per modality while keeping global self-attention and the same FLOP budget as dense models.
  - It acts as a drop-in replacement for dense transformers in Chameleon and Transfusion-style architectures.
- **Results:**
  - Matches or exceeds dense baselines across Chameleon and Transfusion setups while using 40–60% of the pretraining FLOPs and significantly less wall-clock time.
  - Scales well across model sizes and remains competitive with MoE-based baselines at similar or lower compute.
- **Why it matters:**
  - Demonstrates that **structured, modality-aware sparsity** is a practical way to scale multimodal foundation models, preserving rich cross-modal interactions while controlling compute.
  - Provides a clean architectural template for future unified VLMs and multimodal FMs focused on efficiency.

Generated via custom pipeline · 2025-11-19