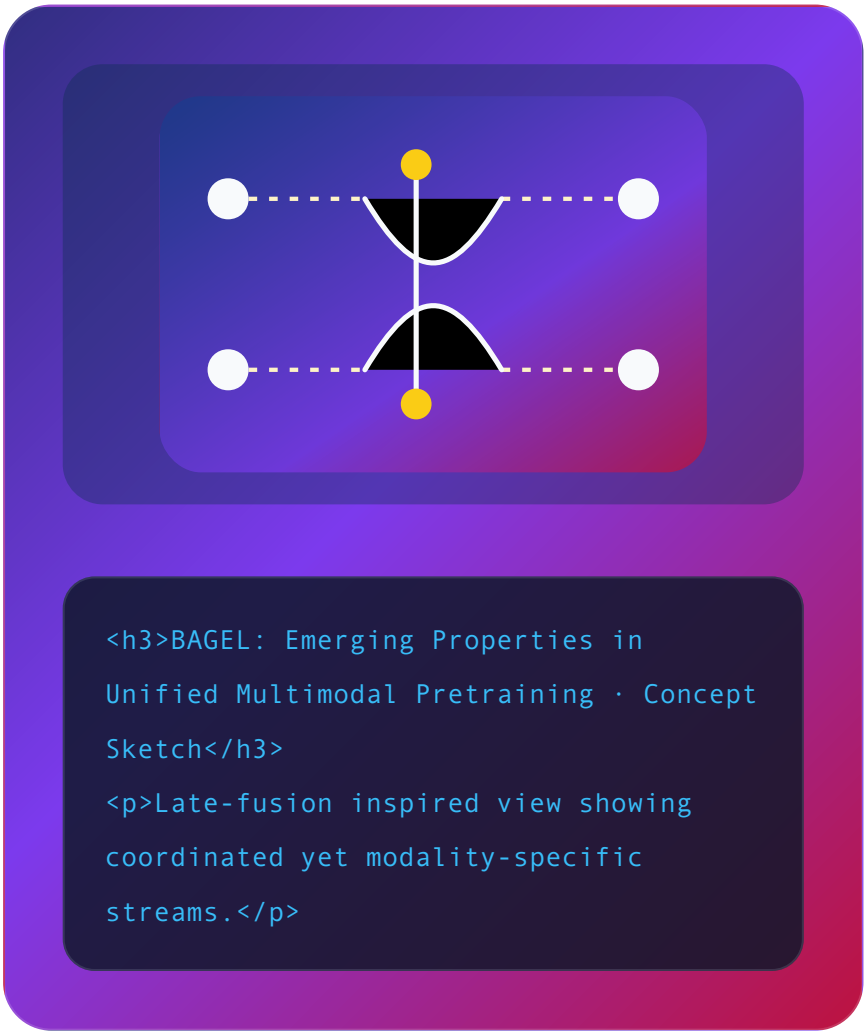

<h3>BAGEL: Emerging Properties in Unified Multimodal Pretraining · Concept Sketch</h3>
Late-fusion inspired view showing coordinated yet modality-specific streams.

BAGEL: Emerging Properties in Unified Multimodal Pretraining

Authors: Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, Haoqi Fan
Year: 2025
Venue: arXiv preprint

---

1. Classification

- **Domain Category:**
  - Vision / VLM / Multimodal FM
  - BAGEL is a unified multimodal foundation model that jointly supports multimodal understanding and generation over text, images, video, and web data.
- **FM Usage Type:**
  - Core FM development + Multimodal FM or cross-modal integration
- **Key Modalities:**
  - Text (natural language).
  - Images (single and multi-image).
  - Video clips.
  - Web and interleaved multimodal content (mixed text–image–video sequences).

---

2. Executive Summary

BAGEL is an open-source unified multimodal foundation model that learns from **trillions of interleaved text, image, video, and web tokens** to support both understanding and generation in a single architecture. The model builds on a decoder-only transformer (initialized from Qwen2.5) and uses a **Mixture-of-Transformer-Experts (MoT)** design with two experts: one specialized for multimodal understanding and the other for multimodal generation, both operating on a shared token sequence through common self-attention. Visual inputs are handled by separate understanding- and generation-oriented encoders (SigLIP-style ViT and a FLUX-derived VAE), while visual outputs are produced via diffusion-style rectified flow conditioned on the transformer states. The authors curate large-scale, reasoning-oriented, interleaved multimodal data and introduce IntelligentBench, a new benchmark suite that better reveals emerging capabilities such as free-form visual manipulation, 3D manipulation, and world navigation. BAGEL outperforms prior open-source unified models on standard multimodal leaderboards and delivers image generation quality competitive with state-of-the-art public generators. For a new grad student, BAGEL illustrates how large-scale interleaved pretraining and unified architectures can close part of the gap between academic models and proprietary systems like GPT-4o.

---

3. Problem Setup and Motivation

- **Scientific / practical problem:**
  - Develop an **open-source unified multimodal model** that can both understand and generate across many

modalities (text, images, video) and tasks (VQA, captioning, editing, navigation, etc.), approaching the capabilities of proprietary systems.

- Study the **emerging properties** that arise when scaling interleaved multimodal pretraining, especially complex compositional reasoning and world-modeling abilities.

- **Why this is hard:**
  - **Data complexity:**
    - High-quality multimodal interleaved data is hard to source, clean, and structure; it must include conversations, instructions, generation tasks, and reasoning traces.
    - Video data is large but crucial for temporal and physical continuity; integrating it at scale is expensive.
  - **Architectural challenges:**
    - Unified models must balance strong language reasoning with high-fidelity visual generation, avoiding bottlenecks between understanding and generation.
    - Naïvely combining autoregressive text and diffusion-style image generation can be compute-intensive and tricky to optimize.
  - **Evaluation gaps:**
    - Standard benchmarks capture basic understanding and generation, but not more advanced capabilities like free-form manipulation, multiview synthesis, or world navigation.
  - **Open-source competitiveness:**
    - Academic models historically trail proprietary systems (GPT-4o, Gemini 2.0) by a wide margin in unified multimodal settings.

4. Data and Modalities

- **Pretraining data:**
  - Large-scale **interleaved multimodal corpus** combining text, images, video, and web content.
  - Data format emphasizes sequences that mix modalities naturally (e.g., conversations with inline images or video frames, web pages with embedded media).
  - Includes **reasoning-oriented content** inspired by DeepSeek-R1: explicit `<think>` segments, chain-of-thought style explanations, and multimodal reasoning traces.
- **Modalities:**
  - **Text:** questions, instructions, descriptions, and dialog.
  - **Images:** high-resolution images for understanding and generation.
  - **Video:** multi-frame clips providing temporal continuity.
  - **Web / interleaved content:** structured pages and documents with embedded media.
- **Preprocessing / representation:**
  - Text is tokenized with the Qwen2.5 tokenizer.
  - Visual understanding uses a **SigLIP2-style ViT encoder** (SigLIP2-so400m/14) with NaViT support for native aspect ratios; outputs image tokens.
  - Visual generation uses a **pretrained FLUX VAE** to convert images to latent tokens (downsampled by 8×, 16 channels), followed by patch embedding to match the transformer's hidden size.

- 2D positional encodings and timestep embeddings are applied to vision tokens for diffusion-style generation.

---

5. Model / Foundation Model

- **Model Type:**
  - Unified multimodal **decoder-only transformer** with a Mixture-of-Transformer-Experts (MoT) architecture: one expert for understanding, one for generation.
- **Is it a new FM or an existing one?**
  - BAGEL is a **new open-source unified multimodal foundation model** that combines MoT, interleaved data, and diffusion-style generation.
- **Key components and innovations:**

Aspect	Details
Backbone	Qwen2.5 decoder-only transformer with RMSNorm, SwiGLU, RoPE, GQA, QK-Norm
Experts	Understanding expert + generation expert; both operate on the same token

Aspect	Details
	sequence via shared self-attention
Visual encoders	SigLIP2-based ViT (understanding) + FLUX VAE (generation) with patch embeddings
Objectives	Next-token prediction for text tokens; rectified-flow diffusion objectives for visual tokens
Parameter scale	7B active parameters (14B total), with MoT structure over shared attention
Data design	Carefully curated, reasoning-oriented, interleaved multimodal data, including video and web

- **Training setup (high level):**
  - Initialize from Qwen2.5 and train on trillions of interleaved multimodal tokens.
  - Use unified decoding: both understanding and generation tasks operate within the same transformer, with expert specialization rather than separate modules.
  - Visual generation uses Rectified Flow, following best practice in large diffusion transformers.
  - Training emphasizes scaling length (context) and variety (modalities, tasks) rather than narrow single-task optimization.

---

6. Multimodal / Integration Aspects (If Applicable)

- **Modalities integrated:**
  - Text, images, and video, with web content as an additional multimodal source.
- **How integration works:**
  - All tokens—text and visual—are placed into a **single long token sequence** and processed with shared self-attention in each transformer block.
  - Two experts (understanding and generation) share this sequence but have distinct parameters; both see the same context, enabling tight coupling between reasoning and generation.
  - Vision tokens come from either the SigLIP-based ViT (for understanding) or the FLUX-based VAE (for generation), but they are mapped into the same hidden space.

- The architecture is **bottleneck-free**: there is no small connector layer compressing context between understanding and generation modules.

- **Why this integration is useful / new capabilities:**
  - Enables **seamless transfer** between understanding and generation: reasoning about input images/videos can directly inform generation within the same context.
  - Long-context self-attention over interleaved sequences supports complex tasks such as world navigation, future-frame prediction, and multi-step visual manipulation.
  - The MoT expert split allows specialization without losing the benefits of a unified transformer.

---

7. Experiments and Results

- **Benchmarks:**
  - Standard multimodal understanding benchmarks (e.g., VQA, captioning, visual reasoning tasks).
  - Image and video generation quality benchmarks (e.g., FID, CLIP-score, human or preference evaluations) compared with open-source generators like SD3 and FLUX.
  - New **IntelligentBench** suite introduced by the authors, focusing on complex multimodal reasoning, free-form manipulation, multiview synthesis, and world navigation.
- **Baselines:**
  - Open-source unified multimodal models and strong VLMs.

    - Public image and video generators (e.g., SD3, FLUX) for generative quality comparisons.
- **Key findings (trends):**
    - BAGEL **outperforms prior open-source unified models** on a broad range of multimodal understanding and generation benchmarks.
    - Image generation quality is competitive with leading public generators, particularly when conditioned on rich prompts and reasoning traces.
    - The model exhibits **emerging abilities**: free-form visual manipulation guided by text and image context, multiview and 3D manipulation, and navigation-style tasks involving world modeling.
    - Scaling interleaved multimodal pretraining reveals a progression: basic understanding/generation → advanced editing and manipulation → long-context reasoning and compositional abilities.

---

8. Strengths, Limitations, and Open Questions

Strengths:

- Provides a **unified, open-source alternative** to proprietary multimodal systems, with strong performance on both understanding and generation.
- Uses a **bottleneck-free MoT architecture** that allows rich interaction between reasoning and generation within a single transformer.
- Emphasizes **interleaved multimodal data** and reasoning-oriented content, which appear crucial for

emergent abilities.

- Introduces IntelligentBench, which better surfaces advanced multimodal reasoning and manipulation capabilities.

Limitations:

- Training requires **huge amounts of data and compute**, limiting reproducibility and accessibility for smaller labs.
- Data curation details—especially for web and video sources—may affect biases and coverage; not all sources are fully public.
- While open-source, deployment may still require substantial GPU resources, particularly for long-context, video-heavy tasks.
- The paper focuses more on performance and qualitative capabilities than on detailed safety, robustness, or bias analyses.

Open questions and future directions:

1. How can we make BAGEL-style unified models more **compute-efficient**, especially for real-time or edge deployments?
2. What are the best ways to **control and align** such powerful multimodal generators, especially for safety-critical or high-stakes tasks?
3. Can similar MoT-based unified architectures be extended to **additional modalities** (e.g., audio waveforms, 3D scenes, sensor data) without major redesign?
4. How should benchmarks evolve to better measure **world-modeling and navigation** capabilities beyond current tasks?

5. What data governance and licensing practices are needed to responsibly scale interleaved multimodal corpora?

---

9. Context and Broader Impact

- **Position in the landscape:**
  - BAGEL is part of a new wave of **unified multimodal foundation models** that aim to close the performance gap with proprietary systems while remaining open-source.
  - It extends integrated transformer-diffusion architectures by adding expert specialization and large-scale interleaved data.
- **Relation to well-known ideas:**
  - Combines ideas from **decoder-only LLMs**, **diffusion transformers**, and **MoT/MoE-style expert architectures** within a single framework.
  - The use of reasoning-oriented pretraining echoes DeepSeek-R1-style chain-of-thought data, but extended to multimodal sequences.
- **Why this paper is a useful reference:**
  - For researchers building VLMs and multimodal FMs, BAGEL provides a blueprint for **scaling unified models** and for designing interleaved training data.
  - It also illustrates the importance of designing new benchmarks (IntelligentBench) that reveal capabilities not captured by traditional tasks.

---

10. Key Takeaways (Bullet Summary)

- **Problem:**
  - Academic unified multimodal models lag far behind proprietary systems in both understanding and generation, especially for complex reasoning and world-modeling tasks.
- **Method / model:**
  - BAGEL is a **7B-active-parameter, MoT-based unified multimodal model** initialized from Qwen2.5, with separate encoders for visual understanding and generation and rectified-flow diffusion for visual outputs.
  - It is trained on large-scale, reasoning-oriented, interleaved multimodal data covering text, images, video, and web content.
- **Results:**
  - Outperforms state-of-the-art open-source unified models on standard multimodal benchmarks and delivers image quality competitive with leading public generators.
  - Exhibits emerging abilities such as free-form visual manipulation, multiview synthesis, and navigation-style tasks, especially as pretraining scale increases.
- **Why it matters:**
  - Shows that **open-source unified multimodal FMs** can approach proprietary systems when trained on rich interleaved data with carefully designed architectures.
  - Highlights the role of **reasoning-oriented multimodal data** and bottleneck-free architectures in enabling complex, world-modeling capabilities.