ByteDance | Seed

Emerging Properties in Unified Multimodal Pretraining

Chaorui Deng[*1], **Deyao Zhu**[*1], **Kunchang Li**[*2‡], **Chenhui Gou**[*3‡], **Feng Li**[*4‡]
Zeyu Wang[5‡], **Shu Zhong**[1], **Weihao Yu**[1], **Xiaonan Nie**[1], **Ziang Song**[1], **Guang Shi**[1§]
Haoqi Fan[*†]

[1]ByteDance Seed, [2]Shenzhen Institutes of Advanced Technology, [3]Monash University
[4]Hong Kong University of Science and Technology, [5]UC Santa Cruz

[*]Equal contribution, [§]Corresponding Author, [†]Project lead

Abstract

Unifying multimodal understanding and generation has shown impressive capabilities in cutting-edge proprietary systems. In this work, we introduce BAGEL, an open-source foundational model that natively supports multimodal understanding and generation. BAGEL is a unified, decoder-only model pretrained on trillions of tokens curated from large-scale interleaved text, image, video, and web data. When scaled with such diverse multimodal interleaved data, BAGEL exhibits emerging capabilities in complex multimodal reasoning. As a result, it significantly outperforms open-source unified models in both multimodal generation and understanding across standard benchmarks, while exhibiting advanced multimodal reasoning abilities such as free-form image manipulation, future frame prediction, 3D manipulation, and world navigation. In the hope of facilitating further opportunities for multimodal research, we share the key findings, pretraining details, data creation protocal, and release our code and checkpoints to the community.

Date: June 13, 2025
Corresponding: shiguang.sg@bytedance.com
Project Page: https://bagel-ai.org/

1 Introduction

The field of unified multimodal understanding and generation has witnessed a surge in interest, with numerous research projects [57, 66, 68, 70, 73, 79, 83, 88] demonstrate promising results in jointly optimizing generation and understanding benchmarks with a crafted unified architecture. While several efforts [1, 11, 70] attempt to scale up their unified models, they are still trained predominantly on image-text paired data from standard image generation and understanding tasks. Recent research [10] has revealed a substantial gap in unified multimodal understanding and generation between academic models and proprietary systems such as GPT-4o and Gemini 2.0, whose underlying techniques remain undisclosed. We argue that the key to close this gap lies in scaling with carefully structured multimodal interleaved data - integrates texts, images, videos and web sources. Our experiments reveal emerging properties as the interleaved multimodal pretraining scales up. Beyond enhancing core multimodal understanding and generation capabilities, the scaling also facilitates complex compositional abilities such as free-form visual manipulation and multimodal generation with long-context reasoning, paving the way for a broad spectrum of advanced functions.

[‡]Work was done during their internship.

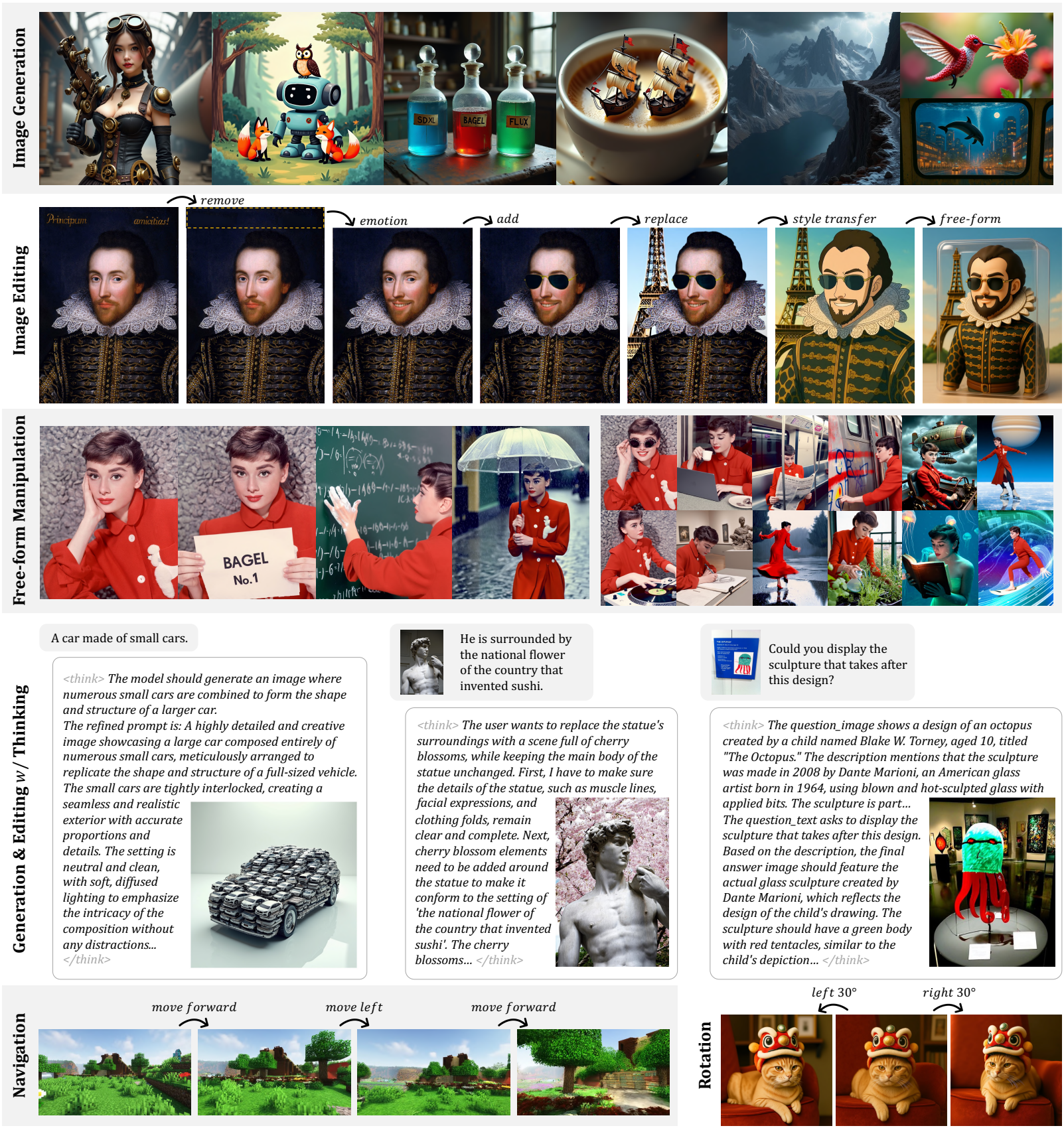

Figure 1 Showcase of the versatile abilities of the **BAGEL** model.

To realize this vision, we established a new protocol for scalable data sourcing, filtering, and construction of high-quality multimodal interleaved data. In addition to the web source, we incorporate video data that naturally provides pixel-level, conceptual, temporal, and physical continuity, which offers exclusive signals essential for acquiring grounded world knowledge at scale. Moreover, our interleaved format inherently includes tasks such as multimodal conversation, text-to-image/video, and image manipulation, enabling seamless integration of diverse generative data. Inspired by DeepSeek-R1 [26], we further enrich the interleaved data with reasoning-oriented content to facilitate multi-modal reasoning, which enables seamless knowledge transfer between understanding and generation processes. As a result, the curated data captures rich world knowledge and nuanced cross-modal interaction content, equipping models with foundational capabilities in in-context prediction, world modeling, and complex multimodal reasoning.

Regarding architecture design, our primary objective is to maximize the capacity of the model without introducing heuristic bottlenecks or task-specific constraints commonly employed in previous models. Following this design philosophy, we adopt a Mixture-of-Transformer-Experts (MoT) architecture that employs selective activation of modality-specific parameters. Unlike some prior approaches [18, 57, 69, 73] that introduce bottleneck connectors between generation and understanding modules, our design enables long-context interaction between multimodal understanding and generation through shared self-attention operations. This bottleneck-free design enables effective scaling of training data and steps, allowing the model's full capacity signals to emerge without being hindered or obscured by architectural constraints.

We present the Scalable Generative Cognitive Model (**BAGEL**), an open-source multimodal foundation model with 7B active parameters (14B total) trained on large-scale interleaved multimodal data. BAGEL outperforms the current top-tier open-source VLMs [4, 12] on standard multimodal-understanding leaderboards, and delivers text-to-image quality that is competitive with leading public generators such as SD3 [19] and FLUX.1-dev [35]. Moreover, BAGEL demonstrates consistently superior qualitative results in classical image-editing scenarios than the leading open-source models. More importantly, it extends to free-form visual manipulation, multiview synthesis, and world navigation, capabilities that constitute "world-modeling" tasks beyond the scope of previous image-editing models. We showcase the qualitative performance in Figure 1.

As BAGEL scales with interleaved multimodal pre-training, we observe a clear emerging pattern: basic multimodal understanding and high-fidelity generation converge first; next, complex editing and free-form visual manipulation abilities surface; finally, long-context reasoning starts to benefit multimodal understanding and generation, suggesting that previously independent atomic skills synergize into compositional reasoning across modalities. These emerging capabilities are not only supported by public benchmarks but are more distinctly revealed in our proposed IntelligentBench, and further verified by qualitative observations. These observations highlight that, while the optimization landscapes for understanding and generation remain partially decoupled, they can be jointly explored via shared self-attention context within a single transformer model, yielding a rich spectrum of capabilities in an open-source system.

2 Model

As illustrated in Figure 2, BAGEL adopts a MoT architecture comprising two transformer experts—one dedicated to multimodal understanding and the other to multimodal generation. Accordingly, the model employs two separate visual encoders: an understanding-oriented encoder and a generation-oriented encoder. The two transformer experts operates on the same token sequence through the shared self-attention operation at every layer. When predicting text tokens, BAGEL follows the Next-Token-Prediction paradigm, adhering to the well-established strengths of autoregressive language models. For visual token prediction, BAGEL adopts the Rectified Flow [19, 41, 45] method following the best practice in the field of visual generation. In the remainder of this section, we share the insights and motivations that shaped these design choices.

2.1 Model Design Space

Typical design choices for unified multi-modal generation and understanding models include:

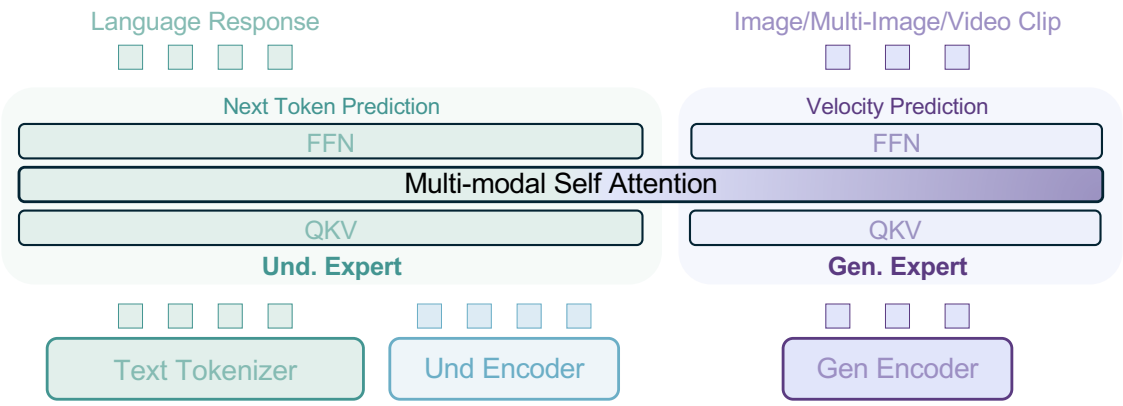

Figure 2 We use **two Transformer experts** to process understanding and generation information, and all tokens do shared multi-modal self attention in each Transformer block. We adopt two distinct encoders to separately capture semantic content and low-level pixel information for image understanding and generation tasks.

Quantized AR. Autoregressive visual generation [11, 48, 59, 70, 79, 83–85, 89] with discrete visual tokenizers [31, 36, 51, 93]. This line of methods leverage the Next-Token-Prediction paradigm for both text and visual token generation, which is straightforward to implement as it can directly utilize existing LLM infrastructures. Unfortunately, the visual generation quality of autoregressive models is empirically inferior to diffusion-based models. Furthermore, inference latency suffers due to the sequential nature of the autoregressive approach.

External Diffuser. LLM backbone combined with an external diffusion module [18, 23, 57, 69, 73]. This design connects pre-trained LLMs/VLMs to diffusion models via lightweight, trainable adapters. Typically, the language backbone autoregressively generates a set of latent tokens as "semantic condition" signals, which are then employed by the diffusion module to generate images. This setup often exhibits rapid convergence with minimal data consumption and may also yield competitive performance [57] on established benchmarks for multi-modal generation and understanding. Its primary drawback, however, is the compression of the LLM context into a relatively small number of latent tokens. This introduces an explicit bottleneck between understanding and generation modules, risking substantial information loss—particularly in long-context multimodal reasoning. Such a constraint might contradict the scaling philosophy of large foundational models.

Integrated Transformer. Unified integration of LLM and diffusion models within a single transformer [40, 50, 66, 102]. Driven by the complementary strengths of autoregressive transformers (powerful understanding/reasoning ability) and diffusion transformers (strong visual generation ability), this approach uses their common model architecture to enable seamless switching between both paradigms. Compared to the External Diffuser solution, it demands substantially higher training compute. Nonetheless, it offers a significant advantage by maintaining a bottleneck-free context throughout all transformer blocks, thereby enabling lossless interaction between the generation and understanding modules and is more amenable to scaling.

In this work, we argue that unified models have the capacity to learn richer multi-modal capabilities from large-scale interleaved multi-modal data—emergent abilities that are not captured by traditional benchmarks. To this end, we choose the bottleneck-free *Integrated Transformer* solution, which we believe to have greater potential in large-scale training settings and may better serve as the foundation model for long-context multimodel reasoning as well as reinforcement learning.

2.2 Architecture

Our backbone model is inherited from an LLM with a decoder-only transformer architecture. We choose Qwen2.5 LLM [92] as the initialization for its superior performance [21] and public availability. It adopts RMSNorm [97] for normalization, SwiGLU [65] for activation, RoPE [67] for positional encoding, and GQA [2] for KV cache reduction. Moreover, we add the QK-Norm [15] in each attention block following the common practice in image/video generation models [19, 35, 63], which is effective in stabilizing the training process.

The visual information is represented from two aspects:

- For *visual understanding*, we leverage a ViT encoder to convert the raw pixels into tokens. We adopt SigLIP2-so400m/14 [75] with a fixed 384-resolution as the initialization of the ViT encoder. Building upon this, we first interpolate the position embedding and set 980×980 as the maximum input size, and then integrate NaViT [16] to enable processing of images at their native aspect ratios. A two-layer MLP connector is adopted to match the feature dimension of the ViT tokens and the LLM hidden states.
- For *visual generation*, we use a pre-trained VAE model from FLUX [35] to convert images from pixel space to latent space and vice versa. The latent representation has a downsample ration of 8 and a latent channel of 16, and is then processed by a 2×2 patch embedding layer to reduce the spatial size and match the hidden dimension of the LLM backbone. The VAE model is frozen during training.

Our framework applies 2D positional encoding to both ViT and VAE tokens prior to their integration into the LLM backbone. For diffusion timestep encoding, we follow [17] and add a timestep embedding directly to the initial hidden states of VAE tokens, instead of using AdaLN as in conventional diffusion transformers [19, 35, 82]. This modification preserves performance while yielding a cleaner architecture. Within the LLM, the text, ViT, and VAE tokens from understanding and generation tasks are interleaved according to the modality structure of input. For tokens belonging to the same sample, we employ a generalized version of the causal attention mechanism. These tokens are first partitioned into multiple consecutive splits, each containing tokens from a single modality (e.g., either text, ViT, or VAE). Tokens in one split may attend to all tokens in preceding splits. Inside each split, we adopt causal attention on text tokens, and keep the bidirectional attention on vision tokens.

2.3 Generalized Causal Attention

During training, an interleaved multimodal generation sample may contain multiple images. For each image, we prepare three sets of visual tokens:

- **Noised VAE tokens**: VAE latents corrupted with diffusion noise, used exclusively for Rectified-Flow training; the MSE loss is computed on this set.
- **Clean VAE tokens**: the original (noise-free) latents, which serve as conditioning when generating subsequent image or text tokens.
- **ViT tokens**: obtained from the SigLIP2 encoder, which help to unify the input format across interleaved generation and understanding data and, empirically, to boost interleaved-generation quality.

For interleaved image or text generation, subsequent image or text tokens may attend to the clean VAE tokens and ViT tokens of preceding images, but *not* to their noised VAE counterparts.

For interleaved multi-image generation, we adopt the diffusion forcing strategy [8], which adds independent noise levels to different images and conditions each image on noisy representations of preceding images. Additionally, to enhance generation consistency, we randomly group consecutive images following [17] and apply full attention within each group. The noise level is the same inside each group.

We implement the generalized causal attention with PyTorch FlexAttention [72], achieving a $\sim 2\times$ speed-up over naive scaled-dot-product attention. During inference, the generalized causal structure allows us to cache key-value (KV) pairs of the generated multimodal context and thus accelerate multimodal decoding. Only the KV pairs of clean VAE tokens and ViT tokens are stored; once an image is fully generated, the corresponding noised VAE tokens in the context are replaced by their clean counterparts. To enable classifier-free guidance [29] in interleaved inference, we randomly drop text, ViT, and clean VAE tokens with probabilities 0.1, 0.5, and 0.1, respectively. An illustration of the generalized casual attention is shown in Figure 15.

2.4 Transformer Design

Following the principle of the Integrated Transformer solution, we compare several transformer variants: the standard Dense Transformer, a Mixture-of-Experts (MoE) transformer, and a Mixture-of-Transformers (MoT) architecture.