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: May 28, 2025

Corresponding: shiguang.sg@bytedance.com

Project Page: https://bagel-ai.org/

arXiv:2505.14683v2 [cs.CV] 23 May 2025

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.

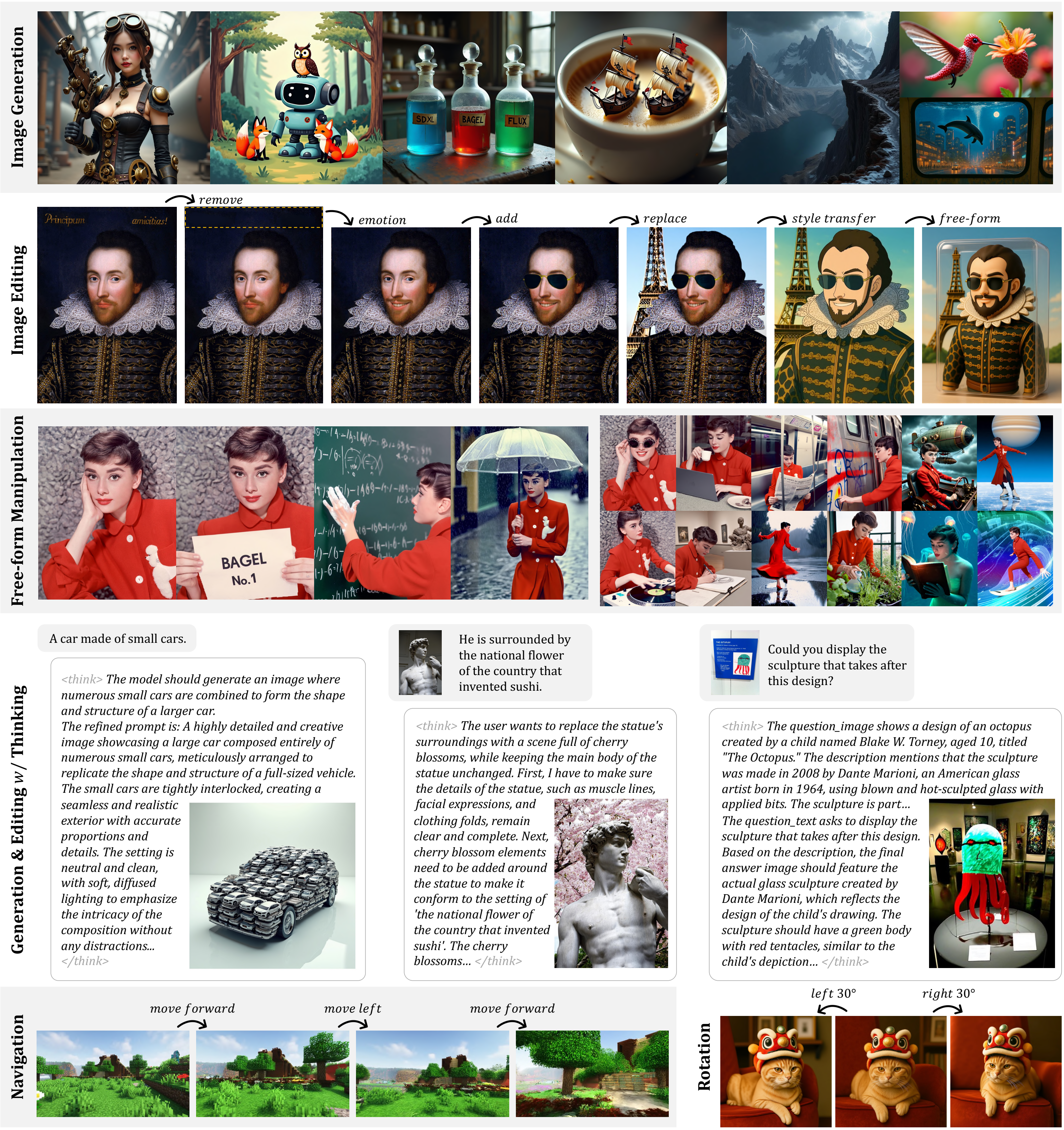


Figure 1 Showcase of the versatile abilities of the **BAGEL** model.

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:

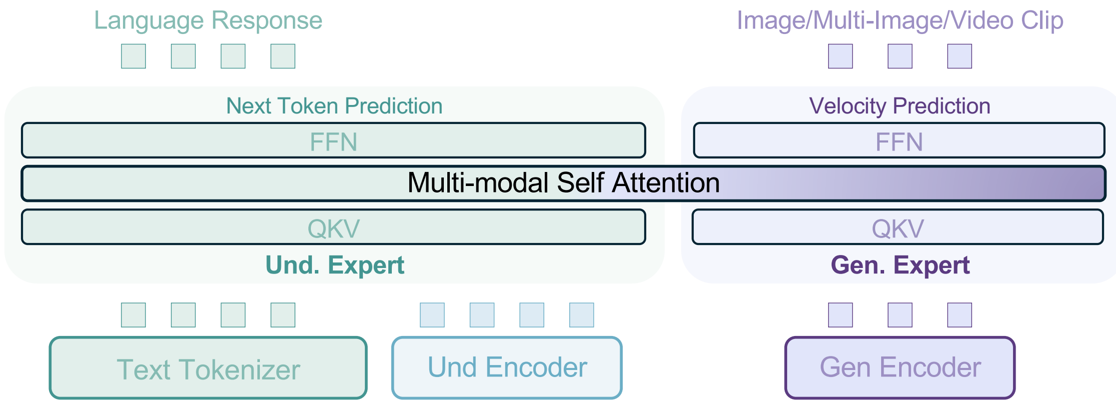


Figure 2 We use **two Transformer experts** to process understanding and generation information, and all tokens do shared multi-modal self attention in each Transformer block. We adopt two distinct encoders to separately capture semantic content and low-level pixel information for image understanding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

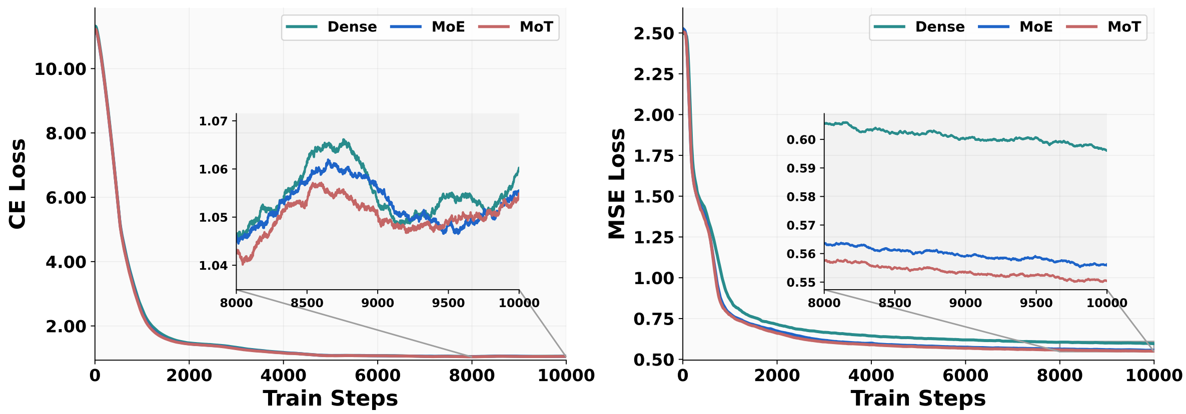


Figure 3 Loss curves of various designs. CE loss and MSE loss are computed on multimodal understanding and generation tasks, respectively. Ablation experiments are carried out on a 1.5B LLM. The sampling ratio for generation and understanding data is set at 4:1.

- **MoE variant**: we duplicate only the feed-forward network (FFN) in each Qwen2.5 LLM block as the initialization of the generation expert.
- **MoT variant**: we duplicate all trainable parameters of Qwen2.5 LLM to create a full-size generation expert. This type of architecture has been adopted by existing works [40, 66].

Both MoE and MoT in our model use hard routing: the newly replicated generation expert exclusively processes VAE tokens, while the original parameters—the understanding expert—handle text and ViT tokens, following the strategy of the Qwen-VL series [4, 77]. Although the MoE and MoT architectures increase the total parameter count by approximately twofold compared to the dense baseline, all three model variants have identical FLOPs during both training and inference.

We conduct a controlled experiment on 1.5B Qwen-2.5 LLM, maintaining identical hyper-parameters and data configurations to isolate the transformer architecture as the sole variable. As illustrated in Figure 3, the MoT variant consistently outperforms both the dense and MoE designs, with the gap being most pronounced on the multimodal generation task. The MSE loss (generation) exhibits a smooth, monotonically decreasing trajectory, where MoT not only converges fastest but also attains the lowest final loss. In contrast, the CE loss (understanding) exhibits greater step-to-step fluctuations—an expected consequence of interleaving heterogeneous data—yet MoT still maintains the best performance in general. These findings highlight the clear advantage of decoupling the parameters devoted to generation from those optimized for understanding, which suggests the two objectives may steer the model toward distinct regions of the parameter space—at least at the 1.5B scale examined here. In short, allocating separate capacity for multimodal understanding and generation can mitigate optimization challenges arising from competing modality-specific learning objectives.

3 Data

As data define the knowledge boundaries of large foundational models, BAGEL is trained on a diverse set of datasets spanning multiple modalities—including language, image, video, and web data—enabling it to perform multimodal reasoning, in-context prediction, physical dynamics modeling, and future frame prediction, all through a unified multimodal interface. In addition to standard vision-language (VLM), text-to-image (T2I), and large-scale language modeling (LLM) datasets, we build new vision-text interleaved datasets from web and video sources to further enhance the model's ability for sequential multimodal reasoning. In Table 1, we summarize the scale and composition of our training data across different modalities. In the following sections, we detail our dataset sources, preparation protocols, and data mixing strategies.

Data Source	# Data (M)	# Tokens (T)
Text Data	400	0.4
Image-Text-Pair Understanding Data	500	0.5
Image-Text-Pair Generation Data	1600	2.6
Interleaved Understanding Data	100	0.5
Interleaved Generation Data: Video	45	0.7
Interleaved Generation Data: Web	20	0.4

Table 1 Data statistics for BAGEL. Since data are randomly sampled during pre-training, the dataset size does not directly correspond to the total number of seen tokens. Multimodal interleaved data is highlight in gray.

3.1 Text Only Data

To maintain the language modeling capabilities of the underlying LLM, we supplement our training corpus with a collection of high-quality text-only data. The data are curated to support broad linguistic coverage and enable strong reasoning and generation abilities across general-purpose text tasks.

3.2 Vision-Text Paired Data

Text-image paired data plays a central role in multimodal learning, providing large-scale visual supervision for both vision-language models (VLMs) [37, 77] and text-to-image (T2I) generation [5, 35, 58, 62]. In our setup, we organize vision-text paired data into two subsets based on their downstream usage: one for VLM pre-training and one for T2I generation.

VLM Image-Text Pairs. We utilize large-scale image-text pairs for VLM training, covering a broad range of visual concepts and primarily sourced from web alt-text and captions. The data have undergone CLIP-based similarity filtering, resolution and aspect ratio constraints, text length checks, and deduplication to ensure quality and diversity. To address long-tail distributions, concept-aware sampling is applied to improve coverage of rare categories. In addition, structured supervision from OCR documents, charts, and grounding annotations is included to enhance the model's capabilities in reading and spatial understanding.

T2I Image-Text Pairs. We incorporate high-quality image-text pairs, as well as minimal synthetic data from existing T2I models [19, 35]. These data feature not only diverse caption styles such as artistic, textual, and surreal captions, but also high-quality images that are filtered for clarity, structural integrity, and semantic diversity. Together, these examples enhance the visual quality and stylistic variety of our T2I training corpus.

3.3 Vision-Text Interleaved Data

While vision-text paired data provides useful supervision, it falls short in supporting complex in-context reasoning involving multiple images and intermediate text. Models trained on such data often struggle to capture visual and semantic relationships across modalities, resulting in less coherent generations. To address these limitations, we incorporate large-scale vision-text interleaved data into training. For improving multimodal understanding, we utilize VLM interleaved datasets. For visual generation, we introduce a unified protocol for constructing vision-text interleaved data by combining diverse sources to support richer multimodal interactions, as detailed below.

3.3.1 Data Source

To comprehensively cover diverse real-world scenarios with scalable data supply, our training corpus integrates two primary sources that provide sufficient knowledge for multimodal reasoning: *video data* and *web data*.

Video data offers rich world knowledge by capturing temporal and spatial dynamics directly from the real world—the largest and most natural simulator. It preserves fine-grained visual details, maintains identity consistency across frames, and models complex motion, making it particularly effective for tasks such as image editing, navigation, and 3D manipulation. We construct our video dataset using publicly available online video resources, as well as two open-source datasets: Koala36M [78], which provides large-scale instructional and interaction-rich content, and MVImgNet2.0 [28], which contains objects captured from varying camera viewpoints to support multi-view spatial understanding.

Filter Type	Description
UI removal	Remove images whose URLs contain substrings such as `icon` or `widget`
Resolution	Require width and height within [150, 20000], and aspect ratio within [1/2, 2]
Image clarity	Remove blurry or low-quality images using a clarity operator
Text density	Discard document-style images with over 100 OCR-detected text tokens
Relevance	Remove redundant or irrelevant images based on CLIP similarity
Doc. trimming	Remove unrelated headers and footers via an LLM
Image quantity	Keep documents with 3–8 images for balanced context

Table 2 Quality filtering rules are applied to web documents, with each filter type accompanied by its specific filtering threshold or method.

Web data captures complex real-world multimodal structures and offers diverse knowledge spanning a wide range of domains. It includes naturally interleaved resources such as illustrated encyclopedic articles, step-by-step visual tutorials, and other richly grounded documents. This interleaved format offers rich supervision for training models to perform multimodal reasoning. We build upon OmniCorpus [39], a large-scale dataset preprocessed from Common Crawl [14], which provides a vast collection of web documents with interleaved text and images. We additionally include open-source image editing datasets as structured interleaved data [3, 22, 32, 80, 87, 100], which teach fine-grained editing behaviors and enhance the model's ability for precise multimodal reasoning and step-by-step generation.

3.3.2 Data Filter

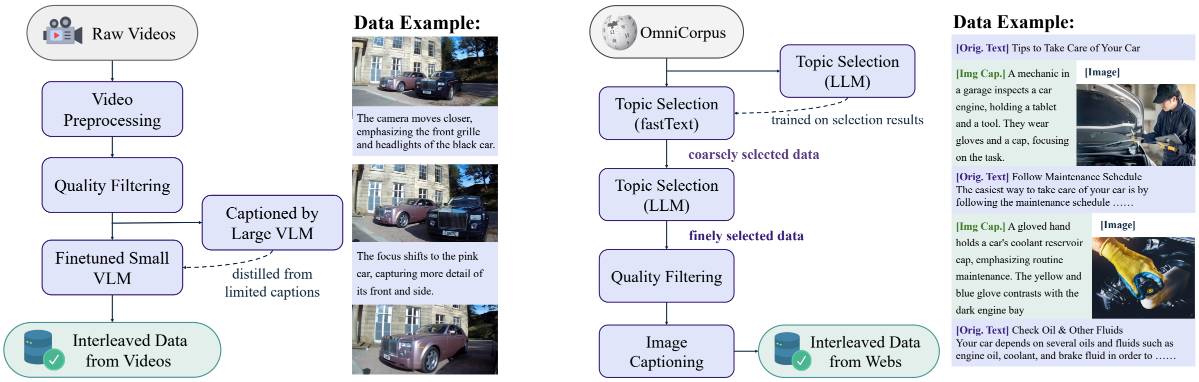
Data Filtering for Video Data. We follow T2V video processing pipelines [63] protocol to preprocess videos into high-quality training clips through temporal splitting, spatial cropping, and quality filtering. Videos are first segmented into short, coherent clips using lightweight shot detection, with related segments optionally merged based on visual similarity. We then remove black borders and overlays such as logos or text using crop detection and frame-level bounding box aggregation. To ensure quality, we filter clips by length, resolution, clarity, and motion stability, and deduplicate using CLIP-based similarity. This process yields a clean and diverse video dataset suitable for multimodal training.

Data Filtering for Web Data. To curate high-quality interleaved data from a large corpus, we design a two-stage filtering pipeline targeting documents such as tutorials, encyclopedic entries, and design content, where text and images exhibit strong semantic alignment. Inspired by DeepSeekMath [64], we first apply a lightweight topic selection process: LLMs are prompted to classify a small subset of documents, and the resulting labels are used to train fastText [34] classifiers for efficient large-scale inference. The selected data are then passed through the LLM classifier again for fine-grained filtering. We adopt the 14B variant of Qwen2.5 models [92] for its balance of performance and efficiency. To further improve data quality, we apply a set of rule-based filters targeting image clarity, relevance, and document structure, as summarized in Table 2.

3.3.3 Data Construction

Interleaved Data from Videos. To construct image-text interleaved sequences from video, we generate textual descriptions of visual changes between consecutive frames—capturing object motion, action transitions, and scene shifts. These inter-frame captions serve as temporal supervision for learning visual dynamics. While large VLMs can produce high-quality change descriptions, their inference cost limits scalability. We instead distill a lightweight captioning model based on Qwen2.5-VL-7B [4], finetuned on a small set of high-quality inter-frame examples. To reduce hallucination, we cap the caption length at 30 tokens. For each video clip, we sample an average of four frames and generate captions for each frame pair, resulting in 45 million temporally grounded interleaved sequences. Figure 4a illustrates the data pipeline along with an example.

Interleaved Data from Webs. To construct high-quality interleaved sequences from web documents, we aim to reduce the difficulty of image generation caused by weak alignment between images, their accompanying text, and surrounding visual context. To provide more localized and relevant cues for each image, we adopt a caption-first strategy: for each image, we generate a concise description using Qwen2.5-VL-7B [4] and insert it directly before the image as a conceptual scaffold. This enables the model to form a conceptual draft of



(a) Data pipeline for interleaved data from videos. **(b)** Data pipeline for interleaved data from webs.

Figure 4 Interleaved data construction pipelines. (a) We construct interleaved video data by preprocessing and filtering raw videos, then generating temporally grounded captions with a small VLM distilled from limited outputs of a large VLM. (b) For web data, we build on OmniCorpus [39] and perform a two-stage topic selection followed by quality filtering and captioning to produce structured sequences. Data examples from both pipelines are shown.

the target image-grounded in both preceding context and the inserted caption—before generating it. By generating the caption to guide what the model should expect in the image, this approach mitigates issues caused by loosely related or ambiguous inputs. Additionally, we rewrite inter-image text segments exceeding 300 tokens using an LLM summarizer to improve contextual density. These steps yield a cleaner and more structured dataset of 20 million interleaved web documents. Data pipeline and examples is shown in Figure 4b.

3.3.4 Reasoning-Augmented Data

Inspired by recent models like O1 [33] and DeepSeek-R1 [26], we leverage long-context Chain-of-Thoughts data for multimodal understanding. Moreover, we hypothesize that introducing a language-based reasoning step before image generation helps clarify visual goals and improve planning. To explore this, we construct 500k reasoning-augmented examples, covering four categories based on the structural relation between input and output: text-to-image generation, free-form image manipulation, and abstract edits.

Text-to-Image generation. We begin by manually crafting a set of brief and ambiguous T2I queries, each paired with simple generation guidance. Using in-context learning, we prompt Qwen2.5-72B [92] to generate additional query-guidance pairs and corresponding detailed prompts, which are then passed to FLUX.1-dev [35] to produce target images. This process yields training triplets of query, reasoning trace (guidance + detailed prompt), and image, enabling models to ground image generation in language-based reasoning.

Free-form image manipulation. We generate reasoning-augmented examples by prompting a VLM with the source image, target image, user query, and a reasoning trace example from DeepSeek-R1 [26]. The R1 example is generated by conditioning on the source and target captions, user query, and a reasoning instruction. The VLM prompt for the reasoning trace generation is demonstrated in Table 9 and Table 10. We sample source and target image pairs primarily from two sources: open-source editing datasets such as OmniEdit [80], and interleaved video data, which provide a rich set of naturally occurring edit scenarios characterized by substantial motion, viewpoint variations, and human interactions while preserving spatial-temporal coherence.

Conceptual Edits. Conceptual edits target cases where image manipulation requires high-level conceptual reasoning rather than simple local pixel modifications, such as transforming an object into a design sketch. For these tasks, we use the web interleaved dataset, sampling candidate image pairs from each sequence and applying a three-stage VLM pipeline to construct high-quality QA examples. First, given a sequence of images, we prompt the VLM to identify a plausible input-output pair. Next, we prompt the model to generate a corresponding textual question based on the selected pair. Finally, we use the VLM to assess the quality of the question and its alignment with the input and output images, filtering out low-quality examples. Accepted examples are then passed to the VLM, prompted with a reasoning trace example from DeepSeek-R1 [26], to produce grounded explanations of the intended transformation, as shown in Table 11. This setup helps the model learn to interpret complex visual goals from diverse textual instructions.

4 Training

As shown in Table 3, we adopt a multi-stage training strategy using a dynamic mixture of the curated data described above—specifically, an Alignment stage for initializing the VLM connector, a Pre-training stage for large-scale pre-training, a Continued Training stage for increased resolution and interleaved data ratio, and a Supervised Fine-tuning stage for high-quality fine-tuning:

- **Stage: Alignment**. In this stage, we align the SigLIP2 ViT encoder with the Qwen2.5 LLM by training only the MLP connector while keeping the vision encoder and the language model frozen. Only image–text pair data are used during this stage to perform image captioning, where each image is resized to a fixed resolution of 378×378 to match the input size of the pre-trained SigLIP2.
- **Stage: Pre-training (PT).** During this stage, we add QK-Norm to the LLM and all model parameters except those of the VAE are trainable. The training corpus comprises 2.5T tokens, consisting of text, image–text pairs, multimodal conversation, web-interleaved, and video-interleaved data. We adopt a native-resolution strategy for both multimodal understanding and generation, with restrictions on the maximum long side and minimum short side of each image.
- **Stage: Continued Training (CT).** Compared with PT, we increase the visual input resolution in the CT stage, which is important for both multimodal generation and understanding performance. We further strategically increase the sampling ratio of interleaved data to emphasize learning cross-modal reasoning, as the model's core understanding and generation capabilities become more stable and reliable. The CT stage consumes approximately 2.6T tokens.
- **Stage: Supervised Fine-tuning (SFT).** In the SFT stage, for multimodal generation we construct a high-quality subset from the image–text-pair dataset and the interleaved-generation dataset. For multimodal understanding, we filter a subset from the LLaVA-OV [37] and Mammoth-VL [27] instruction-tuning data. The total number of training tokens at this stage is 72.7billion.

	Alignment	**PT**	**CT**	**SFT**
Hyperparameters				
Learning rate	1×10^{-3}	1.0×10^{-4}	1.0×10^{-4}	2.5×10^{-5}
LR scheduler	Cosine	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1.0 \times 10^{-15}$)			
Loss weight (CE : MSE)	-	0.25 : 1	0.25 : 1	0.25 : 1
Warm-up steps	250	2500	2500	500
Training steps	5K	200K	100k	15K
EMA ratio	-	0.9999	0.9999	0.995
Sequence length per rank (min, max)	(32K, 36K)	(32K, 36K)	(40K, 45K)	(40K, 45K)
# Training seen tokens	4.9B	2.5T	2.6T	72.7B
Max context window	16K	16k	40k	40k
Gen resolution (min short side, max long side)	-	(256, 512)	(512, 1024)	(512, 1024)
Und resolution (min short side, max long side)	(378, 378)	(224, 980)	(378, 980)	(378, 980)
Diffusion timestep shift	-	1.0	4.0	4.0
Data sampling ratio				
Text	0.0	0.05	0.05	0.05
Image-Text pair (T2I)	0.0	0.6	0.4	0.3
Image-Text pair (I2T)	1.0	0.1	0.1	0.05
Interleaved understanding	0.0	0.1	0.15	0.2
Interleaved generation: video	0.0	0.1	0.15	0.2
Interleaved generation: web	0.0	0.05	0.15	0.2

Table 3 Training recipe of BAGEL. Multimodal interleaved data is highlight in gray .

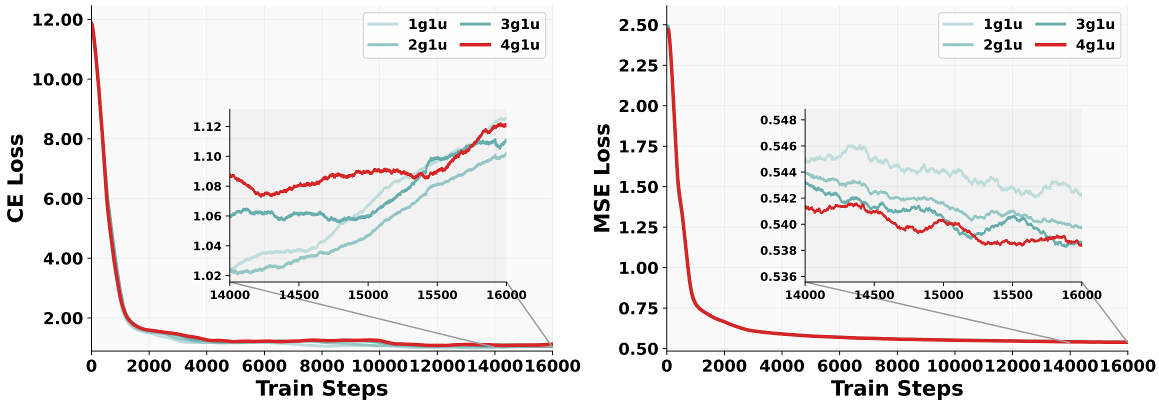


Figure 5 Loss curves of different data ratios. Ablation experiments are carried out on a 1.5B LLM. "1g1u" means that the sampling ratio for generation and understanding data is set at 1:1.

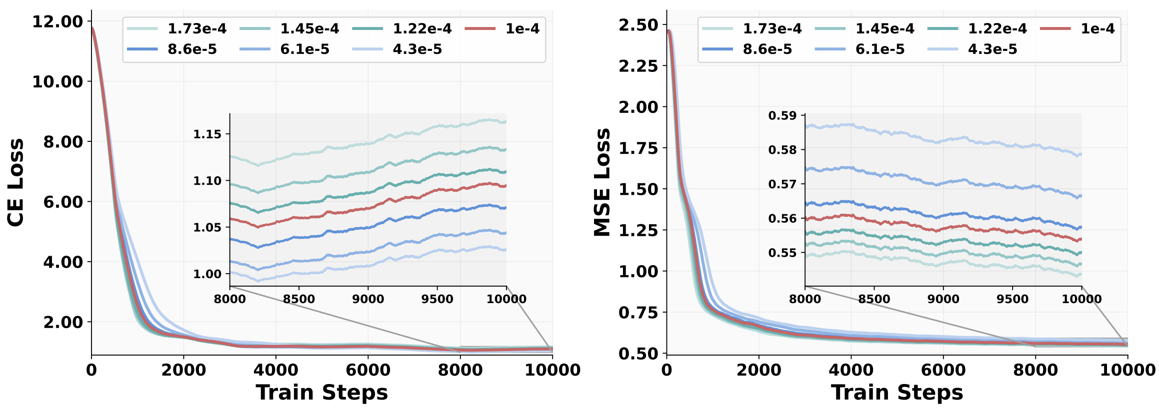


Figure 6 Loss curves of different learning rates. Ablation experiments are carried out on a 1.5B LLM. The sampling ratio for generation and understanding data is set at 1:1.

For all training stages, we use the AdamW [47] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$. Inspired by [52], we set $\epsilon = 1.0 \times 10^{-15}$ to suppress loss spikes. When increasing the resolution for generation, we also increase the diffusion timestep from 1.0 to 4.0 to ensure a proper noise-level distribution. We adopt a constant learning rate for the PT, CT, and SFT stages so that we can easily scale the training data without restarting the training process [30]. To ensure load balance among different ranks, we pack the sequences on each rank into a narrow length range (32K to 36K tokens for Alignment and PT, 40K to 45K tokens for CT and SFT).

Unlike the pre-training of standalone VLMs or T2I models, unified multimodal pre-training requires careful tuning of two key hyper-parameters—the data-sampling ratio and the learning rate—to balance signals from understanding and generation tasks. Below, we describe the empirical insights that guided these choices, which in turn shaped the training protocol summarized in Table 3.

4.1 Data Sampling Ratio

To choose the sampling ratios for each data source during unified pre-training, we conducted a series of controlled studies on the 1.5B Qwen2.5 LLM [92] by adjusting the proportion of multimodal generation data versus multimodal understanding data. As shown in Figure 5, increasing the sampling ration of generation data from 50% ("1g1u") to 80% ("4g1u") steadily reduces the MSE loss, results in a 0.4% absolute reduction—a considerable margin for rectified-flow models in practice. In contrast, the cross-entropy (CE) loss exhibits no consistent pattern across sampling ratios; the largest observed gap, 0.07 at step 14,000 between "4g1u" and "2g1u", has negligible impact on downstream benchmarks. These findings suggest that generation examples should be sampled substantially more often than understanding examples—a heuristic we adopt throughout the training protocol summarized in Table 3.

4.2 Learning Rate

We next carried out a controlled experiment identical to the setup in Section 4.1 except for the learning-rate setup. As shown in Figure 6, the two losses behave oppositely: a larger learning rate makes the MSE loss converge faster, whereas a smaller learning rate benefits the CE loss. To reconcile this trade-off, we assign separate weighting factors to the two objectives, as listed in Table 3.

5 Evaluation

To comprehensively evaluate a unified model, we draw on established benchmarks that target well-defined skills such as multimodal understanding, T2I generation, and classical image editing. Yet for capabilities that demand strong multimodal reasoning and complex task composition, effective evaluation strategies are still lacking. In the following, we first illustrate the available benchmarks used during our evaluation process, and then introduce a new evaluation suite for free-form image manipulation (including conceptual editing), designed to reveal a model's proficiency in multimodal reasoning and complex compositional tasks.

Multimodal understanding. We adopt six widely used benchmarks—MME [20], MMBench (1.0-EN) [46], MM-Vet [95], MMMU [96], MathVista [49], and MMVP [74]. Collectively they offer a concise but comprehensive testbed that spans perception, cognition, and multimodal reasoning, while retaining strong discriminative power for ranking state-of-the-art models.

Text-to-Image generation. We follow [11, 57] and report results on the popular GenEval [25] benchmark. We also adopt the recently proposed WISE benchmark [53], which offers a comprehensive assessment of complex semantic understanding and world-knowledge integration in text-to-image generation. In addition, we include qualitative comparisons with state-of-the-art models as a complement to these automatic evaluation metrics.

Image Editing. We adopt GEdit-Bench [44] as our primary evaluation suite owing to its real-world relevance and diverse set of editing tasks. Built from authentic user requests scraped from the web, GEdit-Bench closely mirrors practical editing needs. Performance is scored automatically with GPT-4.1 [54], and we also supplement these scores with qualitative examples to provide a more nuanced assessment.

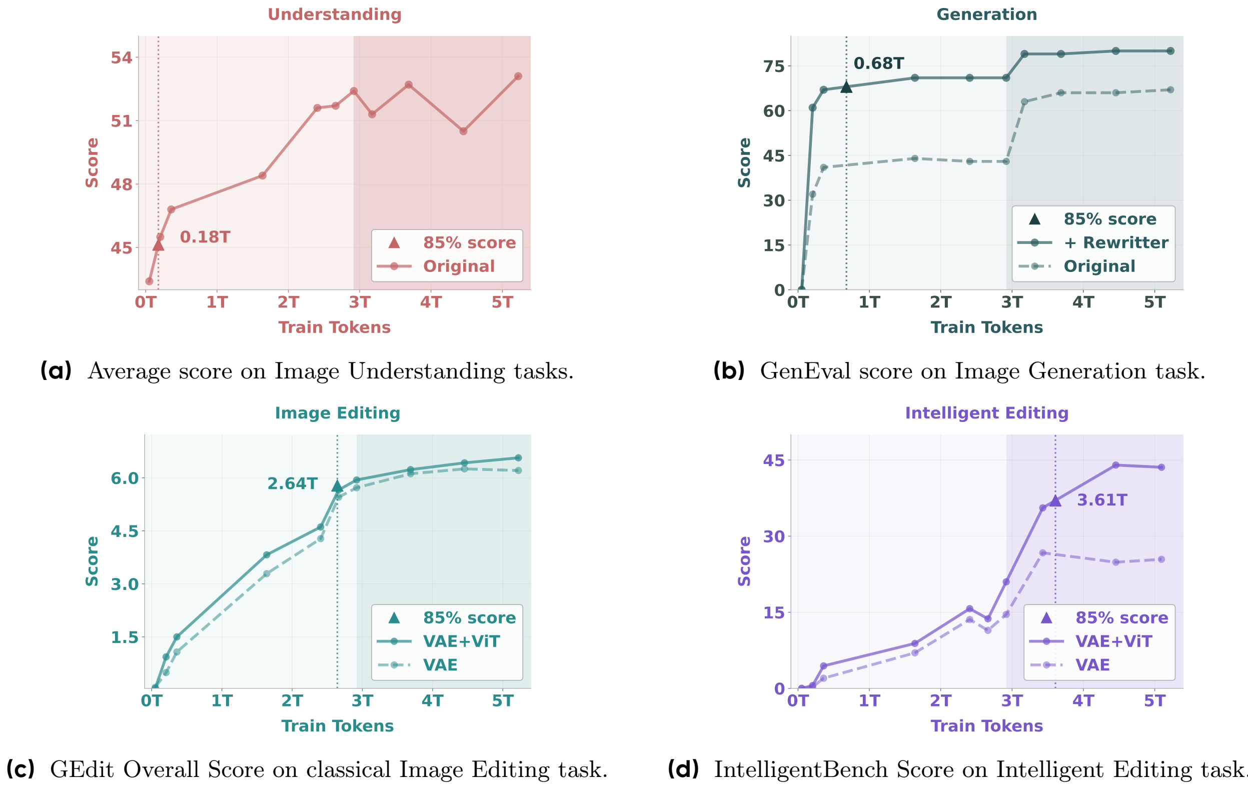
Intelligent Image Editing. We propose *IntelligentBench* as a proxy task for the evaluation of free-form image manipulation ability, which requires complex multimodal reasoning and task composition. The initial release of IntelligentBench comprises 350 examples, each consisting of a question image, question text, and a reference answer image. Evaluation is performed using GPT-4o (version: gpt-4o-2024-11-20), which reviews a complete quadruplet—the question image, question text, reference answer image, and the model-generated image. The evaluation criteria include request fulfillment, visual consistency, and knowledge-grounded creativity, reflecting the benchmark's focus on both task correctness and the depth of reasoning. Each answer is scored on a scale from 0 to 2. The final score of a model is calculated by summing all individual scores and normalizing the total to a 100-point scale. The detailed evaluation prompt is provided in Appendix Table 12. With the help of IntelligentBench, we can evaluate how well the model performs reasoning and integrates world knowledge for image editing. Some showcases and qualitative results on IntelligentBench can be found in Figure 12.

6 Emerging Properties

Emerging properties have been studied extensively in the context of large visual or language models [7, 81]. In this work, situated within the scope of unified multimodal foundational models, we adopt a more focused definition for emerging properties:

An ability is emerging if it is not present in earlier training stages but is present in later pre-trainings.

This qualitative shift, often referred to as a phase transition, denotes a sudden and dramatic change in model behavior that cannot be predicted by extrapolating from training loss curves [81]. Interestingly, we observe the similar phenomenon in unified multimodal scaling, where loss curves do not explicitly signal the emergence of new capabilities. Therefore, we investigate the emergence of model capabilities by evaluating performance across a range of tasks on historical checkpoints. Specifically, we report the average performance on standard VLM benchmarks as a proxy for multimodal understanding, the GenEval score for generation ability, and the GEdit score and IntelligentBench score to assess the model's capability in naive and complex multimodal reasoning, respectively.



(a) Average score on Image Understanding tasks.

(b) GenEval score on Image Generation task.

(c) GEdit Overall Score on classical Image Editing task.

(d) IntelligentBench Score on Intelligent Editing task.

Figure 7 Emerging curves. Pre-training performance curves of BAGEL on different tasks. The lighter region represents the low-resolution pre-training stage, while the darker region indicates the high-resolution CT stage. BAGEL demonstrates consistent performance improvements as the number of training tokens increases. The relationship between performance and training scale can be summarized as follows: **(i) BAGEL continues to improve** across various tasks with more training tokens; **(ii) Different capabilities emerge at different stages**—understanding and generation abilities emerge first, basic editing follows, and intelligent editing emerges last, reflecting the increasing complexity of these tasks. **(iii) Adopting both VAE and ViT features surpasses using VAE features alone in the image editing tasks**, especially in Intelligent Editing, with a noticeable gap. This supports the idea that ViT provides important semantic context to aid generation. Note: The average image understanding score is computed as the mean of the scores from MME-S, MMBench, MMMU, MMVet, MathVista and MMVP. All performance evaluations are conducted with BAGEL's thinking mode disabled.

Interestingly, different tasks demonstrate distinct learning dynamics and saturation behaviors. If we choose the number of seen tokens required to reach 85% of peak performance as an indicator, as noted in Figure 7, we find that conventional understanding and generation benchmarks saturate relatively early: at approximately 0.18T and 0.68T tokens, respectively. In contrast, editing tasks, which require both understanding and generation capabilities, exhibit slower convergence, reaching 85% performance only after 2.64T tokens.

Most notably, the Intelligent Edit task—designed to eliminate naive edit cases and emphasize on complex multimodal reasoning—requires 3.61T tokens to reach 85%, demonstrating a pattern akin to emergent behaviors described in [81]. In this setting, the model shows initially low performance that improves gradually and significantly after the 3T seen tokens. While traditional editing tasks remain largely unaffected by the resolution increase at 3T tokens, Intelligent Editing performance keeps improving significantly—from 15 to 45—tripling in later training stages and underscoring its dependence on unified multimodal reasoning. We further find that understanding ability, particularly visual input, plays a critical role in multimodal reasoning: removing the ViT tokens has minimal impact on GEdit-Bench but causes a 16% drop in Intelligent Edit, highlighting the importance of visual-semantic reasoning in complex editing tasks.

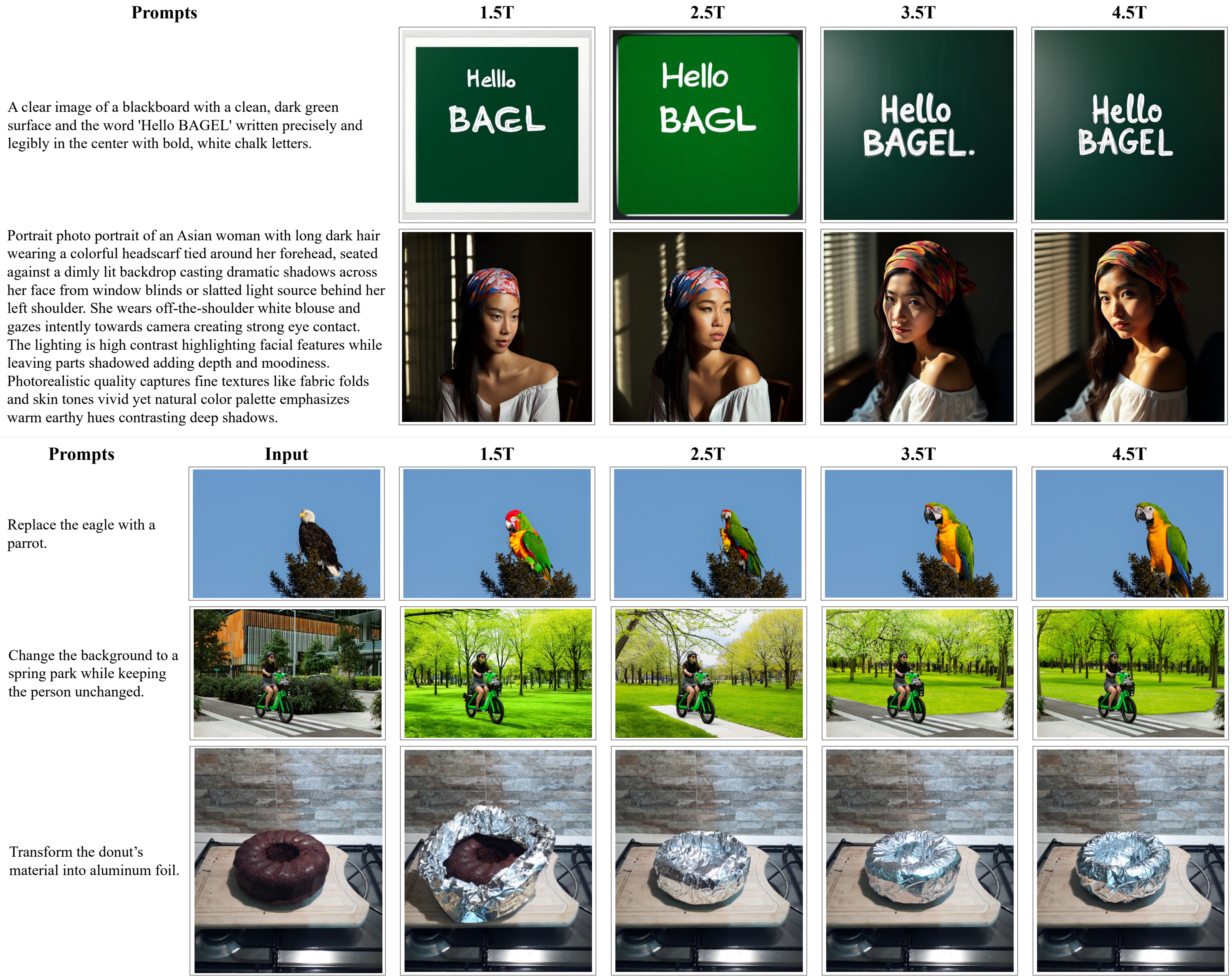


Figure 8 Comparison of models with different amounts of training tokens. We present cases of Text-to-Image generation and image editing.

While evaluation metrics may not linearly capture the model's true capabilities—potentially leading to spurious signs of emergence, albeit unlikely—we further examine qualitative emerging behavior by inspecting generation outputs across different training checkpoints. As illustrated in Figure 8, we observe trends consistent with the performance curves: generation quality is already strong before 1.5T seen tokens , with a small quality improvement after 3.0T seen tokens when trained with higher resolution. For text rendering, the ability to generate correct spell of "hello" and "BAGEL" emerge later—around 1.5T to 4.5T tokens.

The emerging behavior is also observed in the qualitative visualization of Intelligent Editing task in Figure 9. Unlike traditional editing shown in Figure 8, which involves only partial modifications to the input image, Intelligent Editing often requires generating entirely new concept based on multimodal reasoning. Prior to 3.5T tokens, the model tends to reproduce the input image with minimal changes—a fallback strategy when the task is not fully understood. However, after seeing 3.5T tokens, the model begins to demonstrate clear reasoning, producing coherent and semantically appropriate edits, aligning with the emergent behavior seen in Figure 7.

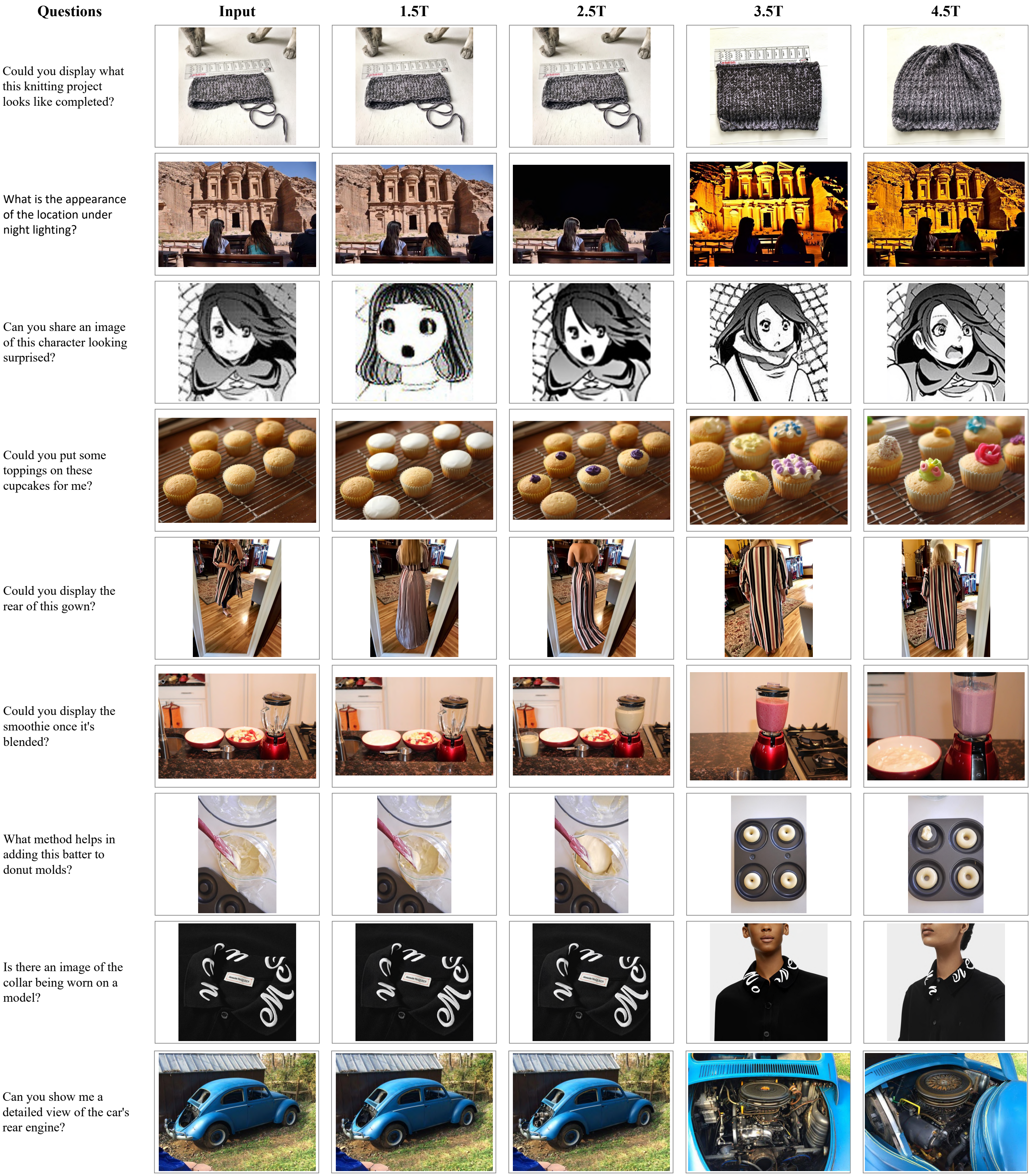


Figure 9 Comparison of models with different amounts of training tokens. We present cases of intelligent editing that requires strong multimodal reasoning abilities.

7 Main Results

In this section, we present both quantitative and qualitative evaluations to examine the diverse multimodal capabilities of BAGEL. We begin with basic abilities on established benchmarks, including image understanding in Section 7.1 and image generation in Section 7.2. We then report performance on existing image editing benchmarks and **IntelligentBench** in Section 7.3. In Section 7.4, we explore generation and editing with explicit reasoning. In this setting, BAGEL is allowed to generate intermediate thinking steps before final outputs. We find that such reasoning significantly enhances performance. Finally, in Section 7.5, we provide qualitative visualizations that showcase BAGEL's world modeling abilities, including world navigation and video generation.

7.1 Image Understanding

Type	Model	# LLM Params	MME-P↑	MME-S↑	MMBench↑	MMMU↑	MM-Vet↑	MathVista↑	MMVP↑
Und. Only	InternVL2 [13]	1.8B	1440	1877	73.2	34.3	44.6	46.4	35.3
	InternVL2.5 [12]	1.8B	-	2138	74.7	43.6	60.8	51.3	-
	Qwen2-VL[77]	1.5B	-	1872	74.9	41.1	49.5	43.0	-
	Qwen2.5-VL[4]	3B	-	2157	79.1	53.1	61.8	62.3	-
	BLIP-3 [90]	4B	-	-	76.8	41.1	-	39.6	-
	LLava-OV [37]	7B	1580	-	80.8	48.8	57.5	63.2	-
	InternVL2 [13]	7B	1648	2210	81.7	49.3	54.2	58.3	51.3
	InternVL2.5 [12]	7B	-	2344	84.6	56.0	62.8	64.4	-
	Qwen2-VL [77]	7B	-	2327	83.0	54.1	62.0	58.2	-
	Qwen2.5-VL[4]	7B	-	2347	83.5	**58.6**	67.1	68.2	-
	Emu3-Chat** [79]	8B	1244	-	58.5	31.6	37.2	-	36.6
	Kimi-VL [71]	2.8B/16B	-	-	-	57.0	66.7	68.7	-
	DeepSeek-VL2 [86]	4.1B/28B	-	-	-	51.1	60.0	62.8	-
Unified	Show-o$_{512}$ [88]	1.3B	1097	-	-	26.7	-	-	-
	Janus [83]	1.5B	1338	-	69.4	30.5	34.3	-	-
	Janus-Pro [11]	1.5B	1444	-	75.5	36.3	39.8	-	-
	BAGEL	1.5B MoT	1610	2183	79.2	43.2	48.2	63.4	54.7
	ILLUME [76]	7B	1445	-	75.1	38.2	37.0	-	-
	VILA-U$^{**}_{256}$ [85]	7B	1336	-	66.6	32.2	27.7	-	22.0
	Chameleon** [70]	7B	-	-	35.7	28.4	8.3	-	0.0
	Janus-Pro [11]	7B	1567	-	79.2	41.0	50.0	-	-
	MetaQuery-XL† [57]	7B	1685	-	83.5	58.6	66.6	-	-
	LlamaFusion** [66]	8B	1604	-	72.1	41.7	-	-	-
	MetaMorph [73]	8B	-	-	75.2	41.8	-	-	48.3
	SEED-X [23]	13B	1457	-	70.1	35.6	43.0	-	-
	TokenFlow-XL [59]	13B	1546	-	68.9	38.7	40.7	-	-
	MUSE-VL [89]	32B	-	-	81.8	50.1	-	55.9	-
	BAGEL	7B MoT	**1687**	**2388**	**85.0**	55.3	**67.2**	**73.1**	**69.3**

Table 4 Comparison with state-of-the-arts on viusal understanding benchmarks. MME-S refers to the summarization of MME-P and MME-C. For MoE models, we report their activate params / total params. †: MetaQuery [57] adopts pre-trained model from Qwen2.5-VL [4] and freezes it during training. **: Partial results are from by MetaMorph [73] or MetaQuery [57].

We extensively benchmark BAGEL against state-of-the-art open-source multimodal models, including both specialized visual understanding and general-purpose unified models. Our evaluation spans a diverse set of public benchmarks to ensure a comprehensive assessment of model capabilities.

The visual understanding results are summarized in Table 4. At a comparable activated parameter size of 7B, BAGEL outperforms existing unified models in understanding tasks. For instance, it achieves significant improvements of 14.3 and 17.1 points over Janus-Pro [11] on MMMU and MM-Vet, respectively. Notably, MetaQuery-XL [57] relies on a frozen, pre-trained Qwen2.5-VL [4] backbone, limiting its adaptability. Moreover, BAGEL delivers superior performance on most of these benchmarks when compared to specialized understanding models like Qwen2.5-VL and InternVL2.5 [12], demonstrating that our MoT design effectively mitigates task conflicts while maintaining strong visual understanding capabilities.

7.2 Image Generation

Type	Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	PixArt-α [9]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [61]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [60]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [79]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [58]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [5]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [19]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
	FLUX.1-dev† [35]	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Unified	Chameleon [70]	-	-	-	-	-	-	0.39
	LWM [42]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X [23]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	TokenFlow-XL [59]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	ILLUME [76]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
	Janus [83]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Transfusion [102]	-	-	-	-	-	-	0.63
	Emu3-Gen†[79]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
	Show-o [88]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
	Janus-Pro-7B [11]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
	MetaQuery-XL† [57]	-	-	-	-	-	-	0.80
	BAGEL	0.99	0.94	0.81	0.88	0.64	0.63	0.82
	BAGEL†	0.98	0.95	0.84	0.95	0.78	0.77	**0.88**

Table 5 Evaluation of text-to-image generation ability on GenEval benchmark. 'Gen. Only' stands for an image generation model, and 'Unified' denotes a model that has both understanding and generation capabilities. † refer to the methods using LLM rewriter.

Type	Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall↑
Gen. Only	SDv1.5 [61]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
	SDXL [58]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
	SD3.5-large [19]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
	PixArt-Alpha [9]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
	playground-v2.5 [38]	0.49	0.58	0.55	0.43	0.48	0.33	0.49
	FLUX.1-dev [35]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified	Janus [83]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
	VILA-U [85]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
	Show-o-512 [88]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
	Janus-Pro-7B [11]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
	Emu3 [79]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
	MetaQuery-XL [57]	0.56	0.55	0.62	0.49	0.63	0.41	0.55
	GPT-4o**	0.81	0.71	0.89	0.83	0.79	0.74	**0.80**
	BAGEL	0.44	0.55	0.68	0.44	0.60	0.39	0.52
	BAGEL *w/* Self-CoT	0.76	0.69	0.75	0.65	0.75	0.58	0.70

Table 6 Comparison of world knowledge reasoning on WISE. WISE examines the complex semantic understanding and world knowledge for T2I generation. 'Gen. Only' stands for an image generation model, and 'Unified' denotes a model that has both understanding and generation capabilities. **: Results of GPT-4o are tested by [91].

We evaluate visual generation performance on two benchmarks: GenEval and WISE. As shown in Table 5, under the same evaluation settings as MetaQuery-XL, BAGEL achieves an **88%** overall score, outperforming both specialized generation models (FLUX-1-dev: 82%, SD3-Medium: 74%) and unified models (Janus-Pro: 80%, MetaQuery-XL: 80%). Even without an LLM rewriter, BAGEL attains 82%, surpassing the previous SOTA unified model, Janus-Pro-7B. On the WISE benchmark, BAGEL exceeds all prior models except the leading private model, GPT-4o. It indicates that BAGEL has strong reasoning ability with world knowledge.

We conduct a qualitative comparison between BAGEL and Janus-Pro 7B, SD3-medium, and GPT-4o. As shown in Figure 10, BAGEL generates significantly higher-quality images than Janus-Pro 7B and also surpasses the widely used specialist text-to-image model SD3-medium. Moreover, it natively supports prompts in both Chinese and English and allows generation at arbitrary aspect ratios.

Figure 10 T2I qualitative comparison. Note that SD3-medium cannot take Chinese prompts so we translate them into English. For GPT-4o, we control the aspect ratio via text prompt. JanusPro can only generates square images.

7.3 Image Editing

Type	Model	GEdit-Bench-EN (Full set)↑			GEdit-Bench-CN (Full set)↑		
		G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Private	Gemini 2.0 [24]	6.73	6.61	6.32	5.43	6.78	5.36
	GPT-4o [55]	**7.85**	**7.62**	**7.53**	**7.67**	**7.56**	**7.30**
Open-source	Instruct-Pix2Pix [6]	3.58	5.49	3.68	-	-	-
	MagicBrush [98]	4.68	5.66	4.52	-	-	-
	AnyEdit [94]	3.18	5.82	3.21	-	-	-
	OmniGen [87]	5.96	5.89	5.06	-	-	-
	Step1X-Edit [43]	7.09	6.76	6.70	7.20	6.87	6.86
	BAGEL	7.36	6.83	6.52	7.34	6.85	6.50

Table 7 Comparison on GEdit-Bench. All metrics are reported as higher-is-better (↑). G_SC, G_PQ, and G_O refer to the metrics evaluated by GPT-4.1.

Type	Model	Score↑
Private	GPT-4o∗∗ [55]	78.9
	Gemini 2.0∗∗ [24]	57.6
Open-source	Step1X-Edit [43]	14.9
	BAGEL	44.9
	BAGEL *w/* Self-CoT	55.3

Table 8 Comparison on IntelligentBench. IntelligentBench examines complex reasoning ability in an image-editing context. ∗∗: Results are reported only on the subset of cases answered (some responses were rejected). GPT-4o answered 318 of 350 questions, while Gemini 2.0 answered 349 questions.

We further evaluate the classical image editing capabilities of BAGEL using the GEdit-Bench [44]. As shown in Table 7, BAGEL achieves results competitive with the current leading specialist image editing model Step1X-Edit [44], and also outperforms Gemini 2.0. Additionally, we report results on our newly proposed IntelligentBench in Table 8, where BAGEL attains a performance of 44.9, significantly surpassing the existing open-source Step1X-Edit model by 30.

We also provide qualitative comparisons across a diverse set of image editing scenarios in Figure 11 and Figure 12, benchmarking BAGEL against Gemini 2.0, GPT-4o, Step1X-Edit, and IC-Edit [99]. As illustrated, BAGEL consistently demonstrates superior performance over Step1X-Edit and IC-Edit, and also exceeds the capabilities of Gemini 2.0. While GPT-4o successfully handles these scenarios, it tends to introduce unintended modifications to the source images, an issue that BAGEL effectively avoids.

7.4 Generation/Editing with Thinking

In this section, we validate the effectiveness of reasoning-augmented generation across various benchmarks from both quantitative and qualitative perspectives.

Generation with thinking. For Text-to-Image task, we evaluate Bagel on WISE with explicit chain-of-thought (CoT) reasoning process before generation. As shown in Table 6, BAGEL with CoT achieves a score of 0.70, surpassing its non-CoT counterpart by 0.18, and also outperforms all existing open-source models by a significant margin (previous SOTA: MetaQuery-XL at 0.55). In addition to the quantitative evaluation, we provide visualizations in Figure 13a, where BAGEL fails to generate correct images when given only a short prompt, but succeeds when using the CoT-based thinking paradigm.

Editing with Thinking. As presented in Table 8, incorporating CoT into BAGEL improves its Intelligent Score from 44.9 to 55.3. This performance gain is primarily attributed to the inclusion of reasoning, which enables the model to leverage world knowledge and provide detailed editing guidance. We further illustrate several representative cases from IntelligentBench in Figure 13b, where the tasks demand general knowledge or multi-step reasoning. In these scenarios, BAGEL demonstrates significantly improved image editing capabilities when guided by the thinking content.

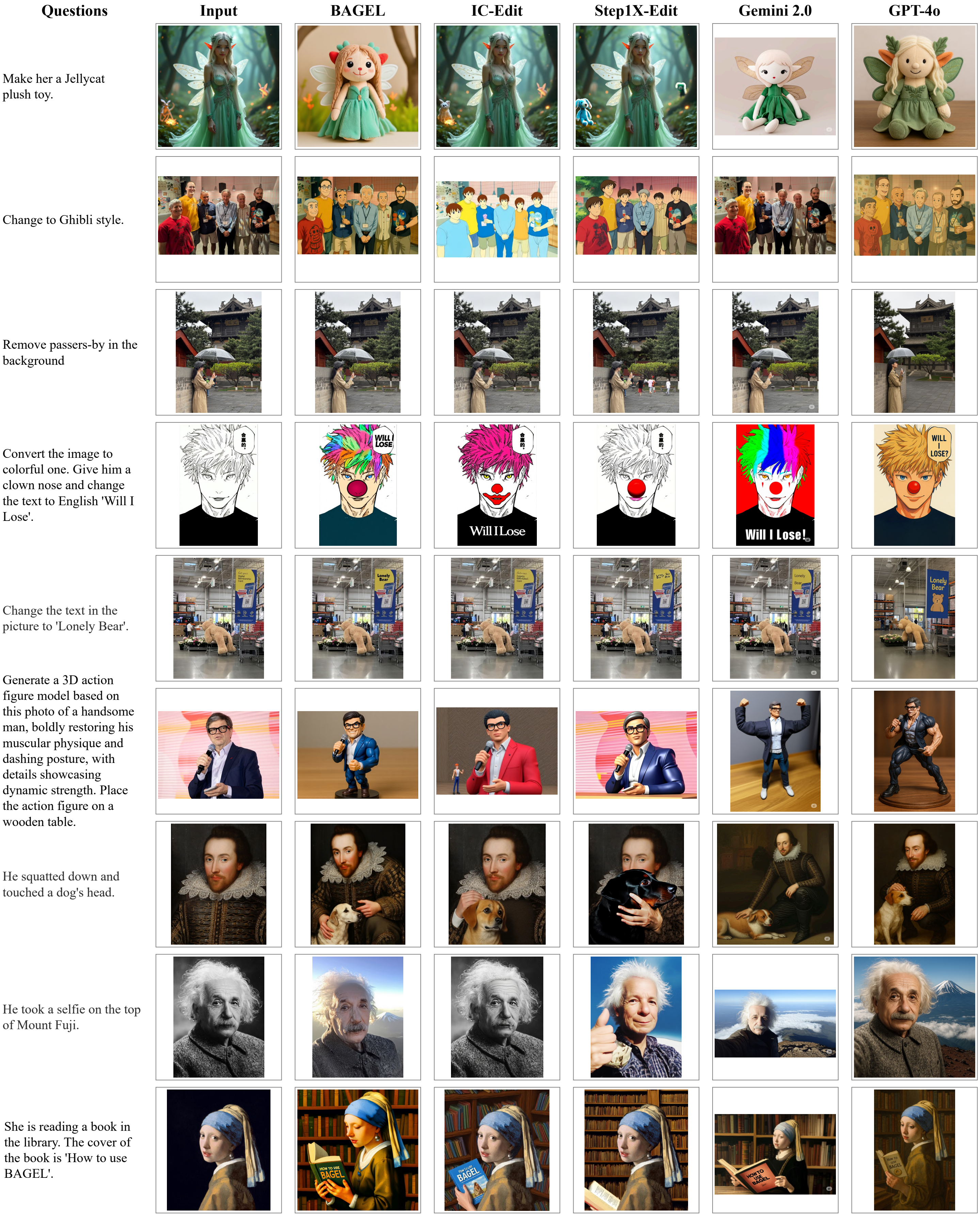


Figure 11 Comparison on editing and manipulation tasks.

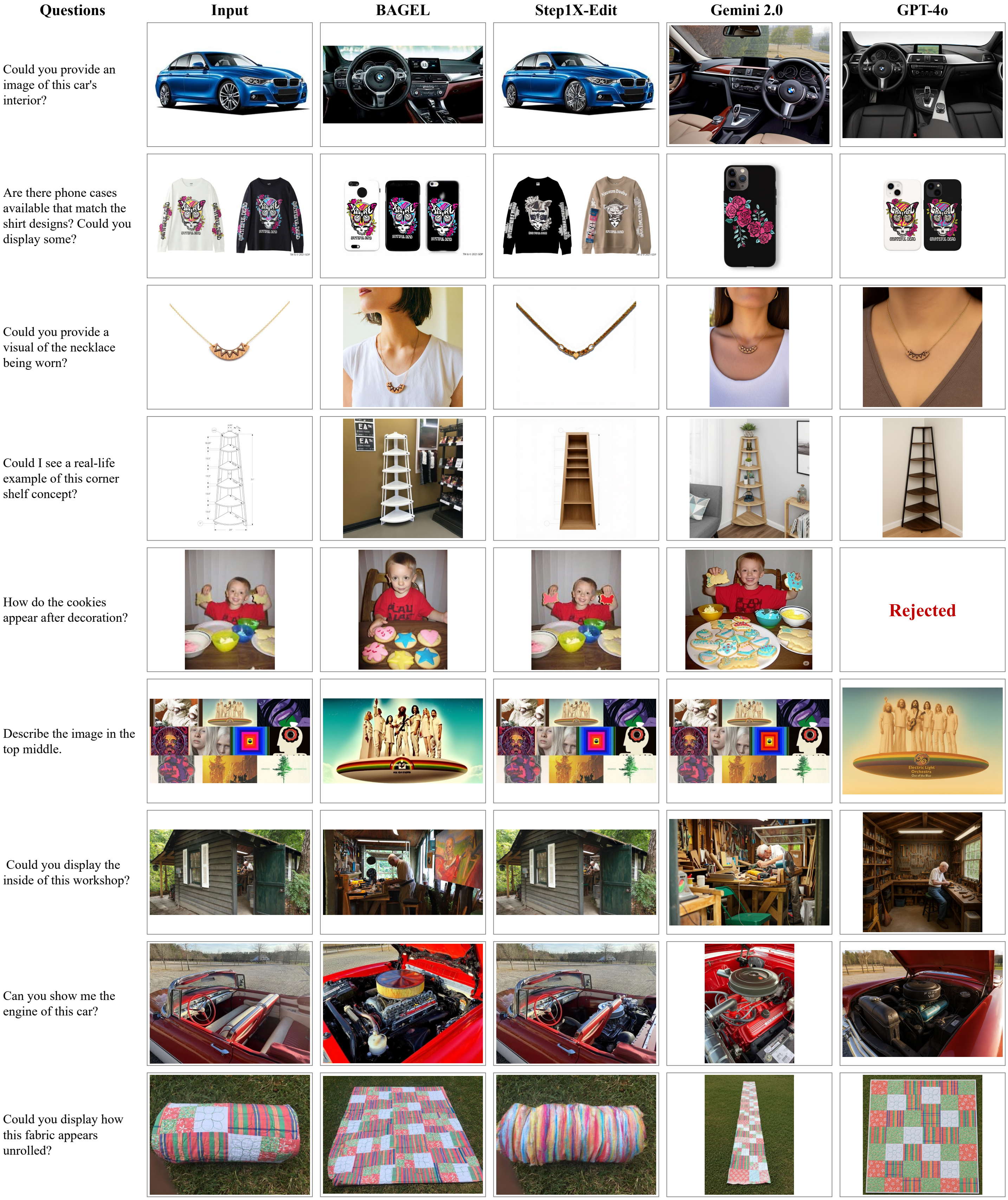
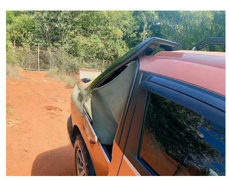
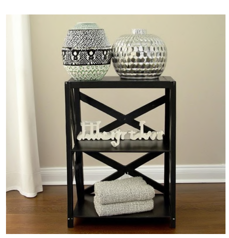


Figure 12 Comparison on IntelligentBench. The results demonstrate that (i) BAGEL achieves performance comparable to Gemini 2.0, effectively handling complex cases that require multi-step reasoning and the incorporation of world knowledge; and (ii) Step1X-Edit fails to address certain instances, often producing outputs that closely resemble the input image, which may be attributed to its limited reasoning capabilities. Note that BAGEL results here are generated in thinking mode.

Prompts	Results	Results *w/* Thinking	
A car made of small cars.		**<think>** The model should generate an image where numerous small cars are combined to form the shape and structure of a larger car. The refined prompt is: A highly detailed and creative image showcasing a large car composed entirely of numerous small cars, meticulously arranged to replicate the shape and structure of a full-sized vehicle. The small cars are tightly interlocked, creating a seamless and realistic exterior with accurate proportions and details. The setting is neutral and clean, with soft, diffused lighting to emphasize the intricacy of the composition without any distractions. The image captures the concept in a photorealistic style, with a focus on precision and craftsmanship, highlighting the innovative use of scale and form. **</think>**	
A cat is twice as large as the dog next to it.		**<think>** The model should generate an image where the cat appears significantly larger than the dog to emphasize the difference in size. The refined prompt is: A highly detailed and realistic image of a large, majestic cat standing next to a smaller dog, with the cat's size being visually emphasized by its towering presence compared to the dog. The cat should have sleek fur, vibrant colors, and a confident, regal posture, while the dog appears smaller, with a friendly and slightly curious expression. The setting is an outdoor environment with soft natural lighting, warm tones, and a gentle atmosphere. The perspective is slightly low-angle, capturing the contrast in size between the two animals in a visually striking composition, with sharp focus on the cat and a subtle depth of field for the background. **</think>**	
A man is standing beside a female brown bear and its cub. Generate an image to show what will likely happen.		**<think>** The model should generate an image of a man in close proximity to the female brown bear and its cub, potentially showing a tense interaction or cautious behavior from the man, as bears are generally protective of their cubs. The refined prompt is: A man stands close to a female brown bear and its cub, depicted in a tense yet natural interaction, with the bear appearing protective and the cub nestled nearby. The scene captures a moment of cautious distance from the man, who exhibits concern or wariness, his posture slightly hunched and alert. The setting is a rugged, forested area with dense greenery and earthy tones, illuminated by soft, natural light filtering through the trees. The mood is tense yet realistic, emphasizing the raw connection between humans and wildlife, with detailed textures on the bear's fur, the man's clothing, and the surrounding foliage. Ultra-realistic, with a focus on lifelike anatomy and environment, capturing the emotional and environmental dynamics. **</think>**	
生成一幅真实水果与微型行星（土星、火星、地球）混合而成的果盘照片		**<think>** 首先，用户需要一幅真实水果与微型行星（土星、火星、地球）混合而成的果盘照片。这意味着需要将三个行星的模型与常见的水果进行融合，形成一个视觉上真实但又富有创意的场景。接下来，我需要考虑如何将这些行星与水果相结合。土星、火星和地球三个行星都具有独特的外观特征，但它们的大小和比例与水果相差悬殊。例如，土星的体积远超一般水果，而火星和地球相对较小。因此，需要将这些行星以比例协调的方式放置在水果盘中，使它们看起来既真实又符合整体构图。水果的选择需要与行星的大小和颜色相匹配。例如，和土星可以与较大的水果（如苹果、橙子）结合，以体现它们的体积。火星和地球可以与较小的水果（如草莓、蓝莓）结合，以保持比例的平衡。此外，还需要考虑光线和阴影，以使行星与水果之间的融合更加自然。光线应从上方或侧面照射，以强调水果的表面纹理和行星的细节。最后，整体构图应具有视觉吸引力，水果和行星的排列要错落有致，避免过于杂乱。背景可以是简单的桌面或果盘，突出前景的水果和行星组合。通过以上思路，最终生成的图像应该是水果与微型行星完美融合的场景，既保持了水果的真实感，又巧妙地展现了微型行星的独特魅力。 **</think>**	

(a) Thinking Helps Generation: Text-to-Image Generation Cases

Questions	Input	Results	Results *w/* Thinking	
Could you transfer the bag from the truck to the SUV?			**<think>** The question involves transferring a bag from a truck to an SUV. The question image shows a truck bed with a large, olive-green bag labeled " keeping calves bags cool with advanced cooling." The bag appears to be designed for storage or transport, likely for items that require temperature control. The question text asks for the bag to be moved to an SUV, suggesting the user wants to visualize the bag in a different context or vehicle. To address this, I would consider the size and shape of the bag, ensuring it fits well in the SUV's cargo area. The SUV in the answer image appears to be a larger vehicle with a spacious cargo area, which should accommodate the bag comfortably. The image would likely depict the bag placed neatly inside the SUV, possibly on the floor or secured in a way that prevents it from shifting during transport. In forming the answer text, I would confirm the transfer and describe the bag's placement in the SUV. The answer image would show the SUV's interior, with the bag visible and properly situated, emphasizing its functionality and utility. This approach ensures the user's request is met clearly and visually. **</think>**	
Could you show me the cabinet with its drawers open so I can see inside?			**<think>** The question text asks for a visualization of the cabinet with its drawers open, allowing the interior to be visible. The question image shows a closed three-drawer cabinet with ornate handles and a metallic finish, placed on a wooden floor. The user likely wants to assess the cabinet's storage capacity or the design of its interior. To address this, I would ensure the answer image depicts the cabinet with all three drawers pulled out, revealing their interiors. The drawers should be open at a standard angle, showing the inside surfaces, which might include lining, compartments, or empty space. The lighting should remain consistent with the original image to maintain continuity, and the cabinet should remain in the same position for clarity. The answer text would confirm that the cabinet is now displayed with its drawers open, fulfilling the user's request for a visual inspection of the inside. This approach ensures the user can evaluate the cabinet's functionality and design. **</think>**	
Could you illustrate how to adorn this shelf with decorative pieces?			**<think>** The question involves removing the tape from the surface of a guitar body that is in the process of being crafted or repaired. The question image shows a guitar body with yellow masking tape applied along the edges, likely to protect the wood during a specific process, such as sanding or painting. The user is asking for the tape to be removed, implying they want to see the surface underneath. To address this, I would focus on ensuring the tape is completely peeled off, revealing the smooth, unobstructed surface of the guitar body. The final image would need to display the guitar body without any tape residue, highlighting its natural wood grain and finish. The workspace in the background, including tools and materials, should remain consistent to maintain context. When describing the answer image, I would note the absence of tape, the clean and polished surface of the guitar body, and any visible signs of craftsmanship, such as the wood grain pattern and the smooth curves of the body. This would fulfill the user's request for a clear view of the guitar body without the tape. **</think>**	

(b) Thinking Helps Generation: Image Editing Cases

Figure 13 Illustration of thinking-aided generation in two tasks. (a) Text-to-image generation. (b) Intelligent editing.

7.5 World Modeling

To improve BAGEL's world modeling ability for long-sequence visual generation, we fine-tune the model by increasing the proportion of video and navigation data in the training recipe. For navigation, we construct our dataset from video interleave sequences, annotating camera trajectories using ParticleSfM [101]. In Figure 14, we demonstrate BAGEL's world modeling capabilities, which include world navigation, rotation, and multi-frame generation.

From the figure, BAGEL exhibits robust world understanding and simulation capabilities. It can follow input instructions to generate a dynamic number of images for tasks like navigating and rotating an input image, or produce multiple images based on a given prompt. Additionally, BAGEL demonstrates strong generalization in world understanding. For instance, while trained solely on real-world street navigation, it seamlessly extends to diverse domains such as ink paintings, cartoons, and video games.

7.6 More Qualitative Results

Performance of BAGEL-1.5B. Figure 16 compares the text-to-image (T2I) and image-editing performance of BAGEL-1.5B—with 1.5 B activated parameters—against JanusPro-7B and Step1X-Edit (12B). Although BAGEL-1.5B is considerably smaller, it surpasses both larger models on both tasks in terms of qualitative comparison. Moreover, the gap between BAGEL-1.5B and BAGEL-7B underscores the gains from model scaling, indicating a greater potential for even larger BAGEL variants.

Failure cases. In Figure 17 we present representative failure cases for BAGEL alongside other state-of-the-art models. Tasks that feature special IP generation, complex textual rendering, intricate human pose generation, or the simultaneous generation of multiple instances remain persistently challenging for contemporary text-to-image systems. For image editing, operations such as swapping object positions or simultaneously modifying a large amount of instances likewise challenge most existing models. In some complex scenarios, both BAGEL and Gemini 2.0 exhibit similar difficulties in adhering precisely to the given instructions. By contrast, GPT-4o delivers the most consistently successful results across all examples. Performance of BAGEL can be simply enhanced by scaling data with additional text-containing images, increasing model capacity, or applying RLHF [56] during the final post-training stage.

8 Conclusion

We presented BAGEL, a unified multimodal understanding and generation model that shows emerging capabilities when scaling up unified pretraining. BAGEL yields top-tier performance on standard multimodal understanding and generation benchmarks, and further distinguish itself with powerful world modeling and reasoning capabilities. In the hope of unlocking further opportunities for multimodal research, we open source BAGEL to the research community.

9 Acknowledgement

We'd like to thank Ziqian Wei, Haoli Chen, Shengyang Xu, Chen Li, Yujia Qin, Yi Lin, Wenhao Huang, Yan Shen, Xiaojun Xiao, Yan Wu, Gang Wu, Guodong Li, Kang Lei, Liang-Wei Tao, Qifan Yang, Bairen Yi, Xiuli Chen, Rui Cao, Yating Wang, Yufeng Zhou, Mingdi Xu, Tingting Zhang, Xuehan Xiong, Tianheng Cheng, Zanbo Wang, Heng Zhang, Yanghua Peng, Faming Wu, Jiashi Feng, Jianfeng Zhang, Xiu Li for their contributions to the BAGEL project.

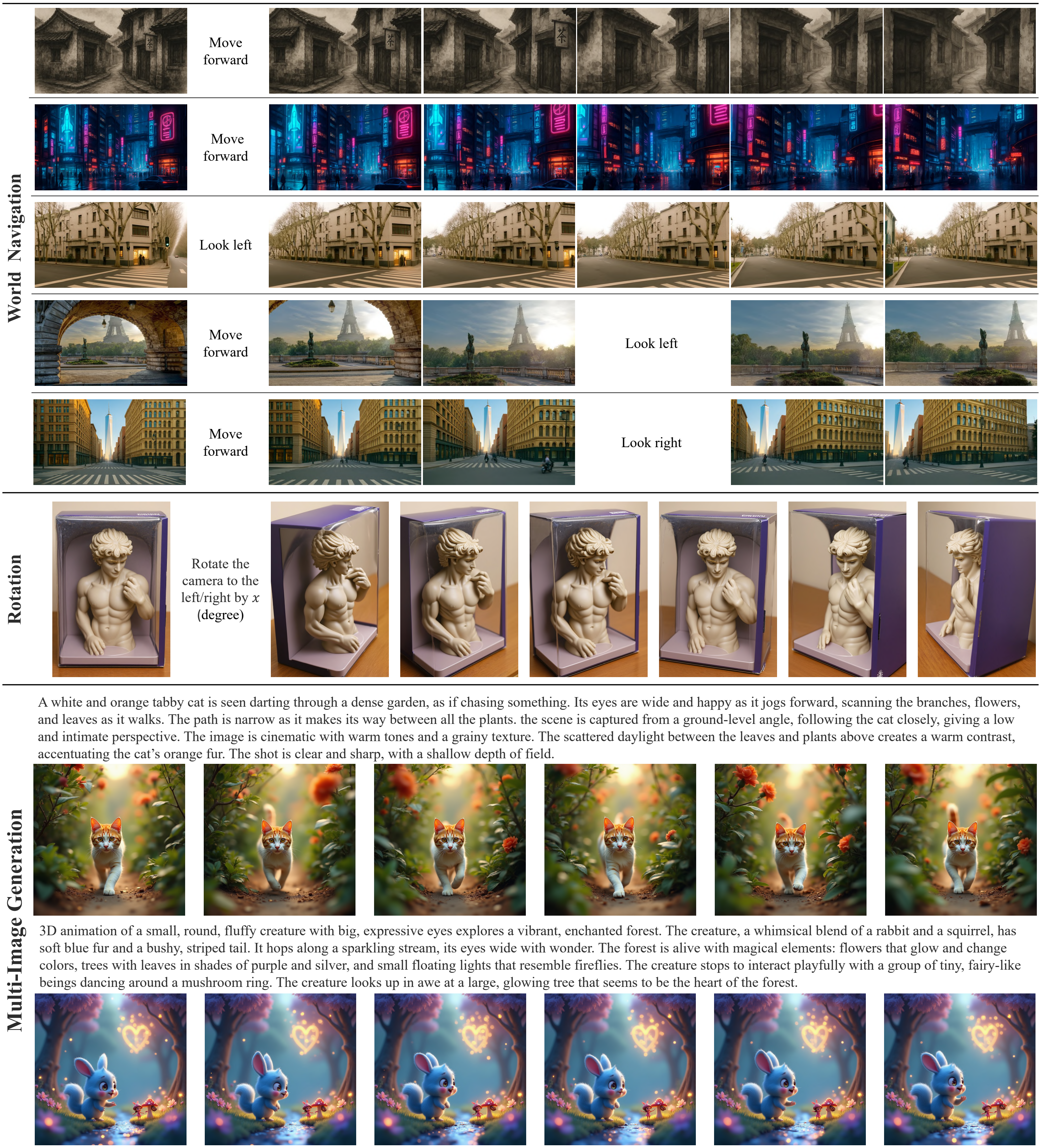


Figure 14 Examples of BAGEL in navigation, rotation, and multi-image generation.

References

[1] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In ICML, 2023.

[2] Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245, 2023.

[3] Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. arXiv preprint arXiv:2412.04280, 2024.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. OpenAI blog, 2023.

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In CVPR, 2023.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In ICCV, 2021.

[8] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In NeurIPS, 2024.

[9] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In ECCV, 2024.

[10] Sixiang Chen, Jinbin Bai, Zhuoran Zhao, Tian Ye, Qingyu Shi, Donghao Zhou, Wenhao Chai, Xin Lin, Jianzong Wu, Chao Tang, Shilin Xu, Tao Zhang, Haobo Yuan, Yikang Zhou, Wei Chow, Linfeng Li, Xiangtai Li, Lei Zhu, and Lu Qi. An empirical study of gpt-4o image generation capabilities. arXiv preprint arXiv:2504.05979, 2025.

[11] Xiaokang Chen, Chengyue Wu, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811, 2025.

[12] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271, 2024.

[13] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences, 2024.

[14] Common Crawl. Common crawl - open repository of web crawl data., 2007. URL https://commoncrawl.org/.

[15] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In ICML, 2023.

[16] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. In NeurIPS, 2023.

[17] Chaorui Deng, Deyao Zhu, Kunchang Li, Shi Guang, and Haoqi Fan. Causal diffusion transformers for generative modeling. arXiv preprint arXiv:2412.12095, 2024.

[18] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In ICLR, 2024.

[19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In ICML, 2024.

[20] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023.

[21] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. arXiv preprint arXiv:2503.01307, 2025.

[22] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. arXiv preprint arXiv:2405.04007, 2024.

[23] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arxiv:2404.14396, 2024.

[24] Google Gemini2. Experiment with gemini 2.0 flash native image generation, 2025. URL https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation/.

[25] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In NeurIPS, 2023.

[26] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[27] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. arXiv preprint arXiv:2412.05237, 2024.

[28] Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. Mvimgnet2. 0: A larger-scale dataset of multi-view images. arXiv preprint arXiv:2412.01430, 2024.

[29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In NeurIPS Workshop on Deep Generative Models and Downstream Applications, 2021.

[30] Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, dahai li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. In COLM, 2024.

[31] Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In ICML, 2023.

[32] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. arXiv preprint arXiv:2404.09990, 2024.

[33] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.

[34] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651, 2016.

[35] Black Forest Labs. Flux, 2024. URL https://github.com/black-forest-labs/flux.

[36] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In CVPR, 2022.

[37] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. TMLR, 2025.

[38] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024.

[39] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. arXiv preprint arXiv:2406.08418, 2024.

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

[41] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In ICLR, 2023.

[42] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. arXiv preprint arxiv:2402.08268, 2024.

[43] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025.

[44] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025.

[45] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In ICLR, 2023.

[46] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In ECCV, 2024.

[47] I Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[48] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In CVPR, pages 26439–26455, 2024.

[49] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In NeurIPS Workshop on Mathematical Reasoning and AI, 2023.

[50] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. arXiv preprint arXiv:2411.07975, 2024.

[51] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. arXiv preprint arXiv:2309.15505, 2023.

[52] Igor Molybog, Peter Albert, Moya Chen, Zachary DeVito, David Esiobu, Naman Goyal, Punit Singh Koura, Sharan Narang, Andrew Poulton, Ruan Silva, et al. A theory on adam instability in large-scale machine learning. arXiv preprint arXiv:2304.09871, 2023.

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

[54] OpenAI. Introducing gpt-4.1 in the api. OpenAI Blog, 2025. URL https://openai.com/index/gpt-4-1/.

[55] OpenAI. Introducing 4o image generation, 2025. URL https://openai.com/index/introducing-4o-image-generation/.

[56] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. ArXiv, abs/2203.02155, 2022. URL https://arxiv.org/abs/2203.02155.

[57] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256, 2025.

[58] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In ICLR, 2024.

[59] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. arXiv preprint arXiv:2412.03069, 2024.

[60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arxiv:2204.06125, 2022.

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684–10695, 2022.

[62] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS, 2022.

[63] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. arXiv preprint arXiv:2504.08685, 2025.

[64] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

[65] Noam Shazeer. Glu variants improve transformer. arXiv preprint arXiv:2002.05202, 2020.

[66] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188, 2024.

[67] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

[68] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222, 2023.

[69] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In CVPR, 2024.

[70] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024.

[71] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report. arXiv preprint arXiv:2504.07491, 2025.

[72] Pytorch Team. Flexattention: The flexibility of pytorch with the performance of flashattention. Pytorch Blog, 2024. URL https://pytorch.org/blog/flexattention/.

[73] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164, 2024.

[74] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In CVPR, pages 9568–9578, 2024.

[75] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

[76] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. arXiv preprint arXiv:2412.06673, 2024.

[77] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

[78] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. arXiv preprint arXiv:2410.08260, 2024.

[79] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. arXiv preprint arxiv:2409.18869, 2024.

[80] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In ICLR, 2024.

[81] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682, 2022.

[82] Peebles William and Saining Xie. Scalable diffusion models with transformers. In ICCV, 2023.

[83] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848, 2024.

[84] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. arXiv preprint arXiv:2412.04332, 2024.

[85] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429, 2024.

[86] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302, 2024.

[87] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340, 2024.

[88] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arxiv:2408.12528, 2024.

[89] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. arXiv preprint arXiv:2411.17762, 2024.

[90] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872, 2024.

[91] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. arXiv preprint arXiv:2504.02782, 2025.

[92] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[93] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In ICLR, 2022.

[94] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. arXiv preprint arXiv:2411.15738, 2024.

[95] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In ICML, 2024.

[96] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In CVPR, 2024.

[97] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In NeurIPS, 2019.

[98] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In NeurIPS, 2023.

[99] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690, 2025.

[100] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In NeurIPS, 2024.

[101] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In ECCV, 2022.

[102] 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.

Appendix

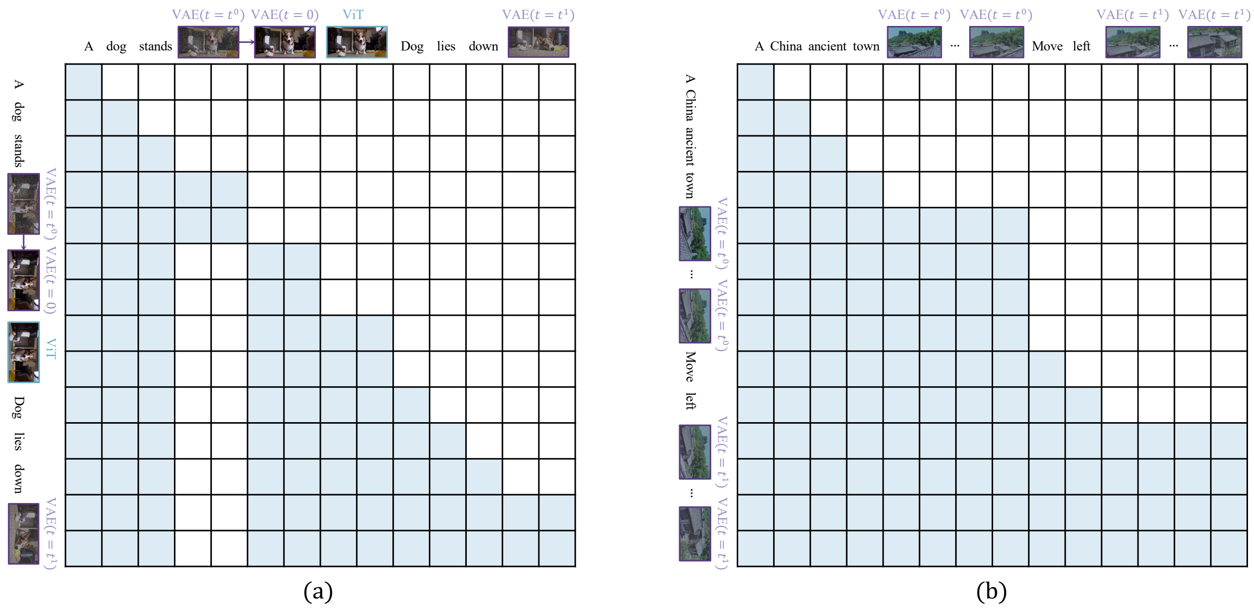


Figure 15 Causal mask in BAGEL during training. VAE and ViT denote VAE features and ViT features, respectively. t is the noise timestep and t=0 means no noise. For each individual image, we apply full attention within its own VAE and ViT features. (a) During interleaved image-text generation, each image attends exclusively to the clean (noise-free) VAE and ViT tokens of preceding images (if present). (b) For interleaved multi-image or video clip generation, we adopt the diffusion forcing strategy [8], conditioning each image on noisy representations of preceding images. Additionally, to enhance generation consistency, we randomly group consecutive images and apply full attention within each group.

Prompts | Janus-Pro-7B | BAGEL-1.5B | BAGEL-7B

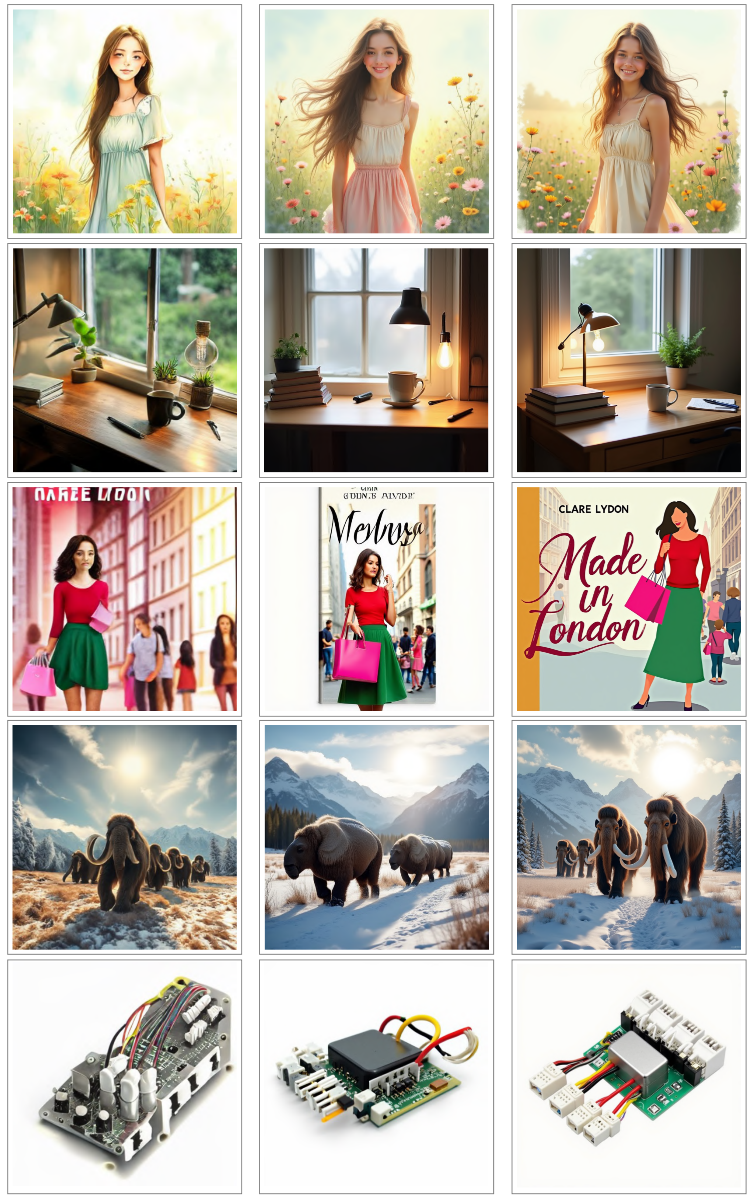
A young woman with long, flowing hair, wearing a vintage sundress, standing in a field of wildflowers. She has a gentle smile, and the sunlight creates a soft, ethereal glow around her. The image is styled in a watercolor painting format, with delicate brushstrokes and pastel colors.

A wooden desk by the window. Its surface is smooth, with a stack of books on one side and a lamp with a soft - glowing bulb on the other. Beside the lamp, there's a half - filled coffee mug and a pen. A small potted plant adds a bit of greenery.

The image is a book cover. It features a woman standing in an urban setting, likely in London, given the theme of the book. The woman is wearing a red top and a green skirt, and she is holding a pink shopping bag, suggesting a relatable, modern-day scenario. In the background, there are other people, including children, which adds to the everyday city life theme. The title of the book, "Made in London," is prominently displayed in a large, elegant, cursive font. The author's name, "CLARE LYDON," is placed above the title. The overall design is vibrant and eye-catching, with a mix of warm and cool tones to create a sense of contrast and appeal.

The image is a book cover. It features a woman standing in an urban setting, likely in London, given the theme of the book. The woman is wearing a red top and a green skirt, and she is holding a pink shopping bag, suggesting a relatable, modern-day scenario. In the background, there are other people, including children, which adds to the everyday city life theme. The title of the book, "Made in London," is prominently displayed in a large, elegant, cursive font. The author's name, "CLARE LYDON," is placed above the title. The overall design is vibrant and eye-catching, with a mix of warm and cool tones to create a sense of contrast and appeal.

The image shows a compact electronic module with multiple wired connections attached via white connectors. The colored wires (red, black, yellow, and others) indicate power, signal, or data transmission. A metallic shield covers a key component, likely for RF or EMI protection. The board features multiple ports, allowing extensive wired interfacing. Two small push buttons and surface-mount components are visible. The design suggests it is a wired communication or control module, possibly for IoT, telemetry, or embedded systems, relying on physical connections for data and power transmission.



Questions | Input | Step1X-Edit | BAGEL-1.5B | BAGEL-7B

Could a heart-shaped crystal be incorporated into this bracelet design?

Could you provide a picture of these shoes being worn?

Could you show me this cat sitting upright?

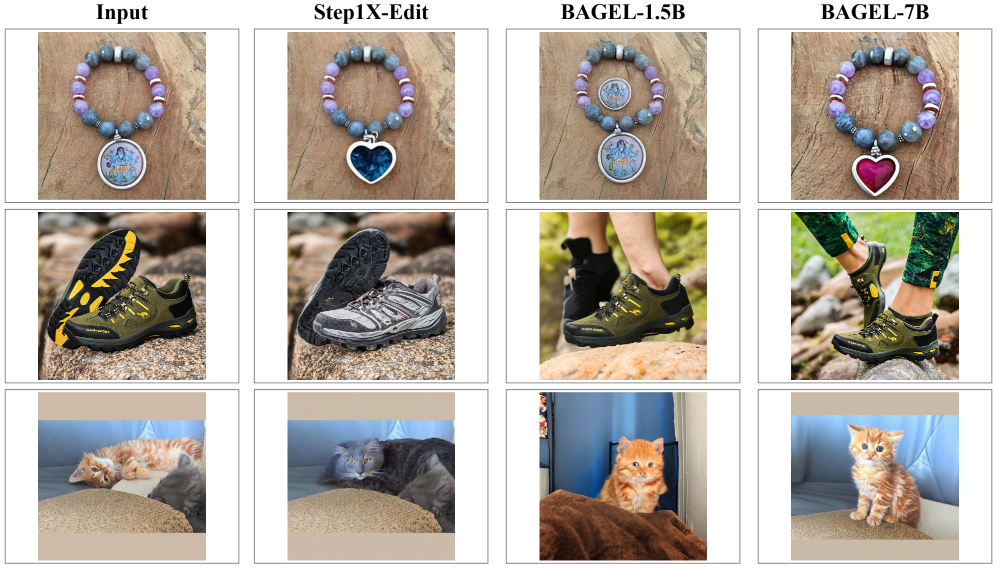


Figure 16 Effect of model scaling: larger models demonstrate better prompt adherence and produce higher-quality images.

Prompts	BAGEL	Doubao	Gemini 2.0	GPT-4o

一张狗仔队偷拍风格的照片，爱因斯坦 (Einstein) 匆忙穿过美国购物中心的停车场，他带着惊讶的表情瞥了一眼，试图避免被拍到。他手里拿着几个铿亮的购物袋，里面装满了奢侈品。他的外套在风中飘动，其中一个包在摇摆，好像他正在大步前进。模糊的背景与汽车和发光的商场入口，以强调运动。相机发出的闪光部分过度曝光了图像，给人一种混乱的小报感。[**16: 9**]

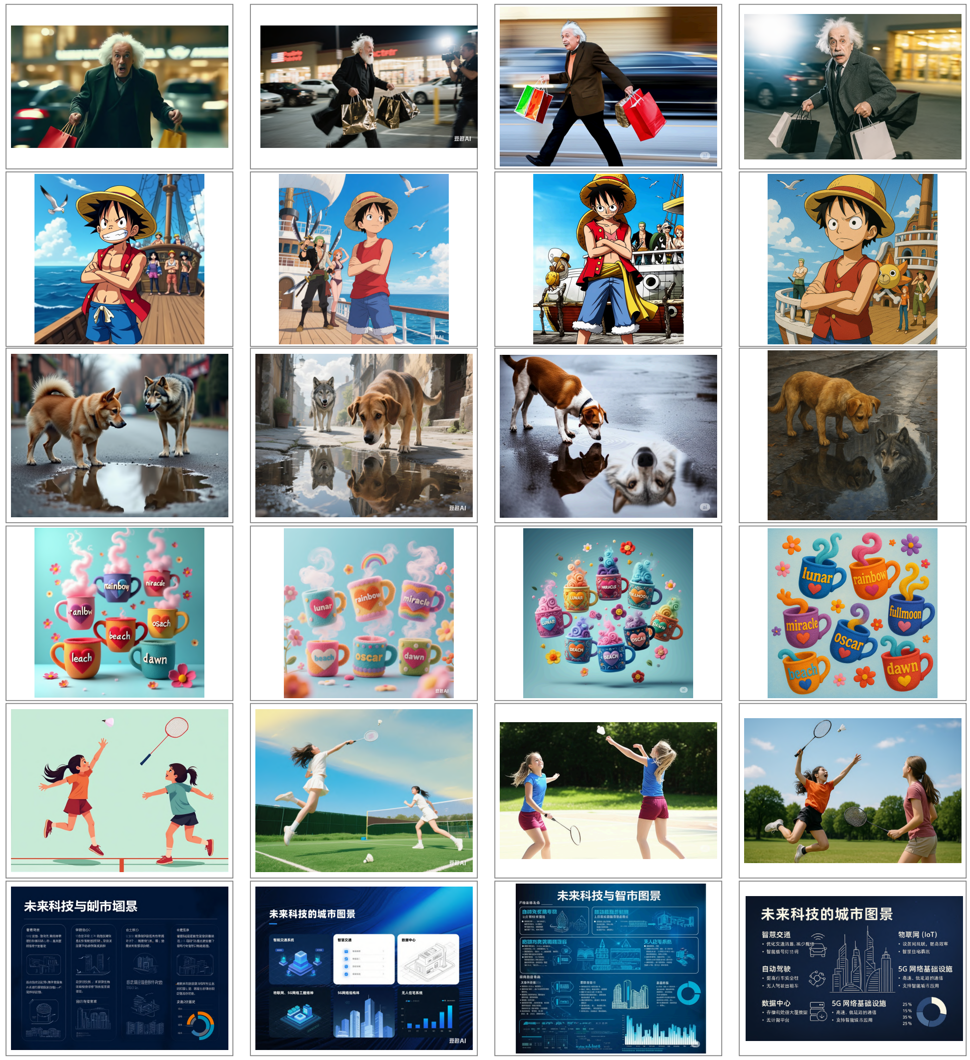
A young Monkey D. Luffy, the main character from One Piece, standing on the deck of the Thousand Sunny. He has a determined expression with his straw hat tilted slightly to one side. Luffy is wearing his iconic red vest, blue shorts, and straw hat. His arms are crossed, and he looks out over the ocean with a sense of adventure. The background shows the Thousand Sunny ship with its crew members in the background, including Zoro, Nami, and Usopp. The sea is calm, with a clear blue sky and a few seagulls flying overhead. The overall style is vibrant and dynamic, capturing the adventurous spirit of the series. [**1: 1**]

A dog looking into a puddle of water on a street, but its reflection is a wolf. [**4: 3**]

A captivating and vibrant image, 3D render, featuring seven colorful, ornate felt mugs, each adorned with a heart and displaying bold text representing the days of the week: "lunar", "rainbow", "miracle", "fullmoon", "beach", "oscar", "dawn". These lively mugs are filled with whimsical felt smoke, and they elegantly float in a dreamy, enchanting atmosphere. The diverse array of floating flowers adds depth and dimension to the scene, while the soft baby blue background harmoniously complements the design. fashion, illustration, typography, 3d render, painting. [**1: 1**]

Two girls are playing badminton. One of them leaps high, ready to hit the shuttlecock. [**4: 3**]

一张视觉精美、信息丰富的长方形幻灯片PPT，主题为"未来科技与智能城市"。风格、科技感十足，整体排版清晰、专业，结构完整。幻灯片顶部是用中文写成的大标题"未来科技的城市图景"，使用无衬线字体，醒目现代页面。其中包含多个区域内容，展示有关智能交通系统、自动驾驶、物联网（物联网）、5G网络基础设施等信息，每个部分都有简洁的中文段落说明和要点列表，如"智慧交通"、"数据中心"、"无人住宅系统"等关键词以加粗或高亮的方式承载。页面中简洁清晰的图标、线条风格的插图、未来城市的建筑草图图、科技设备的概念图。右下角是一个中文标注的数据图表（如柱状图或环形图）。背景为深蓝或渐变色调，带有抽象纹理。整体高对比，布局平衡网格，图文并茂，幻灯片应为完整内容，不能留白或模板感。[**4: 3**]



Prompts	Input	BAGEL	Doubao	Gemini 2.0	GPT-4o

A seagull is standing on the back of the eagle with its wings open.

Swap the position of the blue bottle and the red bottle.

Deblur the image.

Change to 3D animated style.

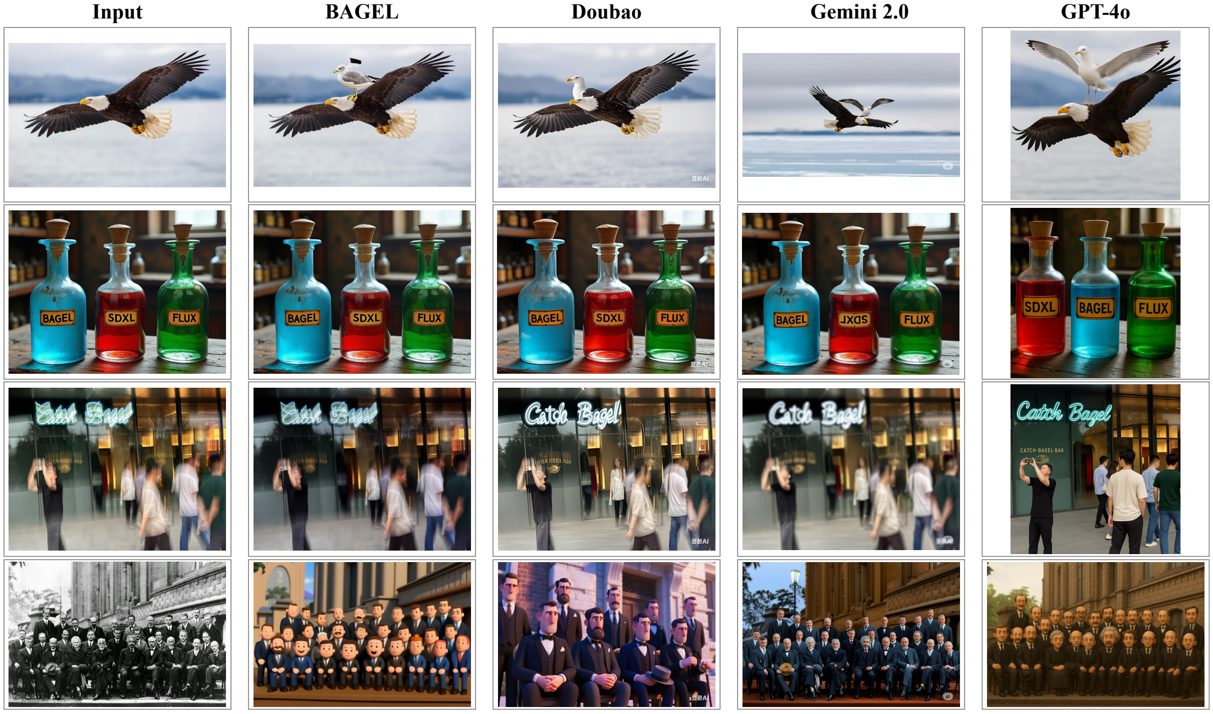


Figure 17 Failure cases. Tasks involving certain IP, complicated text, counterfactual scenes, object swapping, and deblurring pose challenges for BAGEL and other models. In contrast, GPT-4o demonstrates more consistent success in these scenarios.

```
###[System Role Instruction]
```

You have the following information:
1. question image: [Place or reference the question image here]
2. question text: [Place the text of the question here]
3. answer image: [Place or reference the final answer image here]

Your task is NOT to output the final answer or the image. Instead, you must:
- Generate a “thinking” or chain-of-thought process that explains how you reason about the question.
- Provide the reasoning/analysis that leads to the answer image.
- The reasoning/analysis should include what should be changed in the answer image compared to the question image and what should be kept the same.
- The reasoning should highlight that the input image structure and layout should be kept the same.

Below is an example of how your output should look. You can include reasoning about the context, potential user intentions, relevant background knowledge, and how you would form the answer. The length of outputs should be **around or shorter than 60 tokens**.

Example Output:
The user wants to change the background from a sunny garden to a snowy setting. The structure and layout of the pink unicorn with bubble details and sunglasses should remain unchanged. Only the environment needs modification: replacing green grass with snow and surrounding greenery with frosted, snow-covered plants while maintaining lighting coherence.

Table 9 The prompt to generate reasoning trace for Free-form image manipulation from edit data.

###[System Role Instruction]
You have the following information:
1. question image: [Place or reference the question image here]
2. question text: [Place the text of the question here]
3. answer image: [Place or reference the final answer image here]

Your task is NOT to output the final answer or the image. Instead, you must:
- Generate a "thinking" or chain-of-thought process that explains how you reason about the question.
- Provide the reasoning/analysis that leads to the answer image.
- The reasoning/analysis should include what should be changed in the answer image compared to the question image and what should be kept the same.

Below is an example of how your output should look. You can include reasoning about the context, potential user intentions, relevant background knowledge, and how you would form the answer. The length of outputs should be **around or shorter than 60 tokens**.

Example Output:
First, I notice the cat's determined action in pressing a button. To adjust for the answer, the focus shifts to expressing excitement or eagerness. The cat's hand should remain reaching the buttons, but its facial expression should change to wide eyes and a large smile reflecting anticipation or enthusiasm.

Table 10 The prompt to generate reasoning trace for Free-form image manipulation from video interleaved data.

```
###[System Role Instruction]
```

You have the following information:
1. question image: [Place or reference the question image here]
2. question text: [Place the text of the question here]
3. answer text: [Place the final answer text here]
4. answer image: [Place or reference the final answer image here]

Your task is NOT to output the final answer or the image. Instead, you must:
- Generate a detailed "thinking" or chain-of-thought process that explains how you reason about the question.
- Do NOT include the final answer text in your output.
- Provide only the reasoning/analysis that leads to the final answer and the answer image (even though you will not reveal the final answer itself).
- The reasoning/analysis should include some description of the answer image to help the answer-image-generation.

Below is an example of how your output should look. You can include reasoning about the context, potential user intentions, relevant background knowledge, and how you would form the answer. The length of outputs should be **around or shorter than 200 tokens**

Example Output:
First, I notice the user wants to see a vehicle displayed while it's moving. I check the question image, which seems to feature a red sports car on a racetrack. The question text, 'Can you display the vehicle while it's moving?', suggests they want a visual depiction of a car in motion. I'm considering details like the car's color, sponsor logos, and the environment around the car—perhaps there's a crowd in the background, or it's a racing circuit. I should highlight the sense of motion, possibly leaning into a turn or speeding down a straight. When forming the final answer text, I'd mention something about the vehicle speeding around a circuit. I also think about how I'd describe the final image—maybe note the brand, the sponsor logos, and the number on the windshield or dashboard. Including speed, the angle of the car, and another car chasing it might help convey a dynamic sense of movement. Lastly, I recall that the user specifically asked to 'display the vehicle while it's moving,' so I'd ensure the image description references motion, leaning into a turn, and the impression of high velocity. This approach should fulfill their request.

Table 11 The prompt to generate reasoning trace for conceptual editing.

Question	***Input***	***Reference GT***	***Answer***
Could you provide an image of this car's interior?			

`###Human:`

You are given a question, the corresponding question image, a human answered image, and the model-generated (AS) answer image.

Your task is to evaluate whether the AS answers the question based on the following criteria:

Must Exact Fulfillment of Request: The answer image must fulfill the request made in the question. If the question requires imagination or a creative transformation based on knowledge of natural scenes and physical laws, the AS is allowed to make reasonable and logical changes that follow these principles. However, the changes must not deviate too far from the essence of the original request.

Must Satisfy Completeness: Every element requested in the question must be reasonably present and completed in the answer image. Missing elements should be noted, but some degree of creative interpretation is acceptable as long as the request is overall fulfilled.

Must No Visual Errors: The answer image must not contain major visual errors such as proportion issues, blurriness, or logical inconsistencies. Minor imperfections that do not affect the overall quality or coherence are acceptable, but significant visual errors should be avoided.

Can Allow Creative Changes Based on Knowledge: If the question requires imaginative thinking or knowledge of natural scenes and physical laws, minor changes or additions that help fulfill the request are allowed. These changes should align with the natural world, physical principles, or the context of the question. However, large or inconsistent changes that break the scene's logic or introduce factual inaccuracies are not acceptable.

The human answered image is just an example answer for your reference to understand how to answer this question. The AS does not need to be the same as the human answered image. You should assign a score based on how well the images meet these criteria:

0: The AS can't be used for answering this question based on previous criteria. Compared with the AS, the human answered image is significantly better.

1: The AS can answer the question, but is worse than the human answered image in terms of quality.

2: The AS can answer the question with similar or better quality than the human answered image.

{ "score" : int, "reason_of_score" : "Detailed explanation of the reasoning for the score." }

Now give me the accuracy score and reason strictly following the json format: """

`###Assistant:`

{'reason': 'The AS correctly provides the interior image of a BMW car, which matches the request based on the exterior shown in the question image. The interior features, including the steering wheel with the BMW logo, dashboard, and overall design, are excellently captured without significant visual errors or missing elements. Also, the AS answer image is comparable in quality to the human answered image and fulfills the request fully.',
'score': 2}

Table 12 Example of IntelligentBench Evaluation: Please note that the "Human answered image" is simply a label for the GPT-4o's reference to the ground truth (GT) image, rather than an actual image generated by a human.