BAGEL

Emerging Properties in Unified Multimodal Pretraining

Deng et al., ByteDance Seed

Blurb

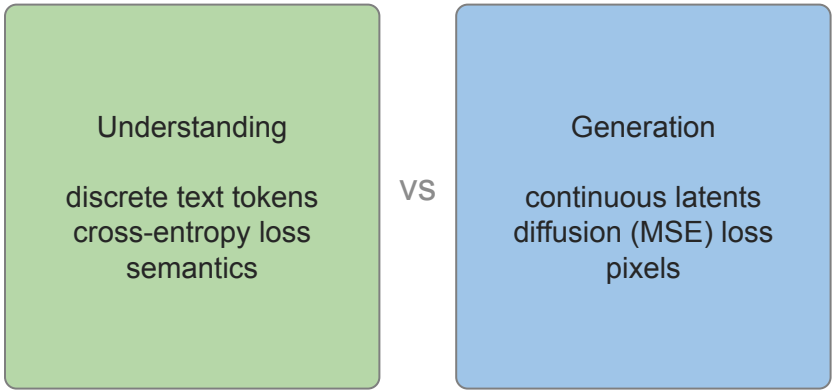
“When scaled with such diverse multimodal interleaved data, BAGEL exhibits emerging capabilities in complex multimodal reasoning”



Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

The Problem

- Goal: one model that can both see and draw
- Open unified models lag far behind GPT-4o / Gemini 2.0
- **Why care:**
  - Does understanding the world make you better at drawing it?
- **Why it's hard:**
  - The two tasks want different things



Can one set of weights do both?

Design Space

Quantized AR

e.g. Janus, Chameleon, Emu3

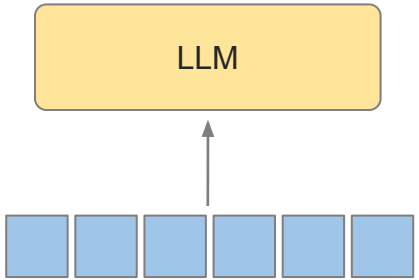
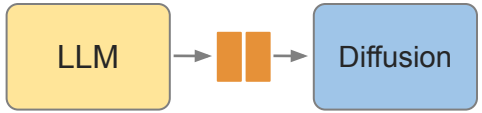


image → discrete codes

Simple, reuses LLM infra
but worse images, slow decoding

External Diffuser

e.g. MetaQuery, MetaMorph

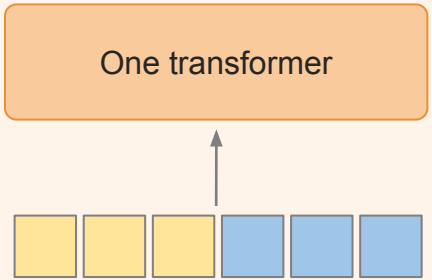


a few latent tokens
= bottleneck

Cheap and quick to train
but a bottleneck between the two

Integrated Transformer

e.g. Transfusion, LlamaFusion



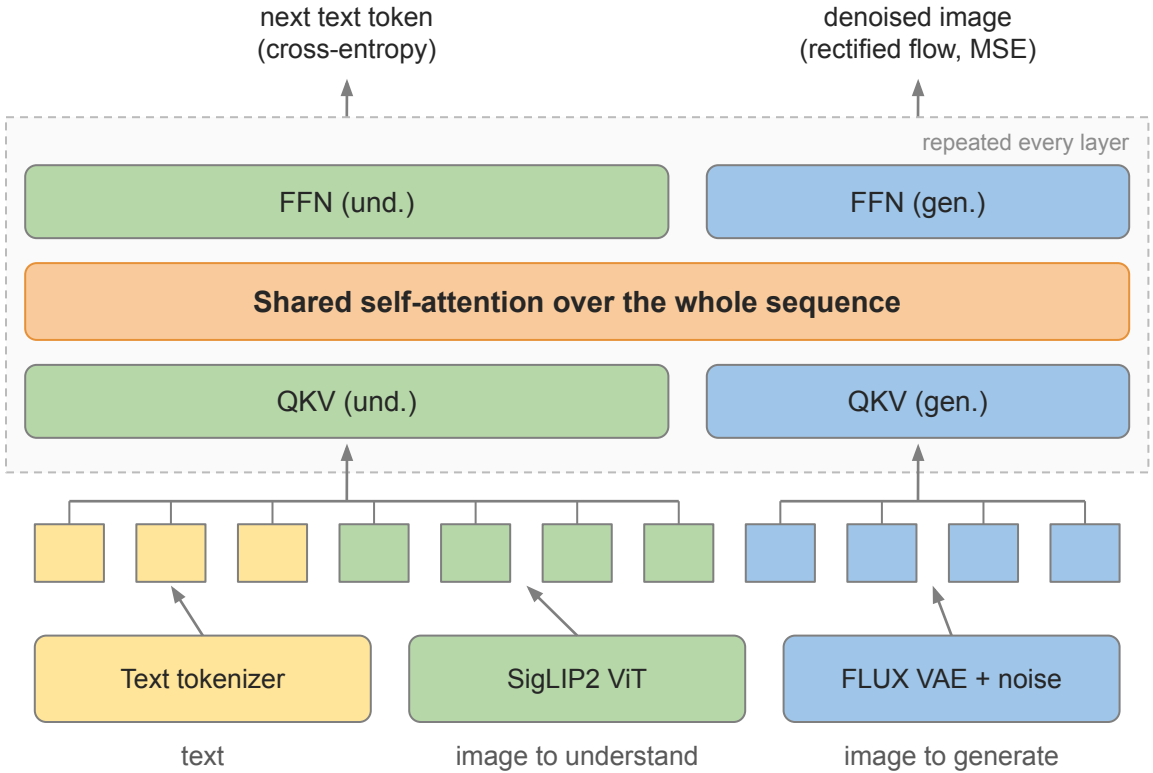
text + image tokens, same context

No bottleneck, scales well
but a lot more compute

The Model

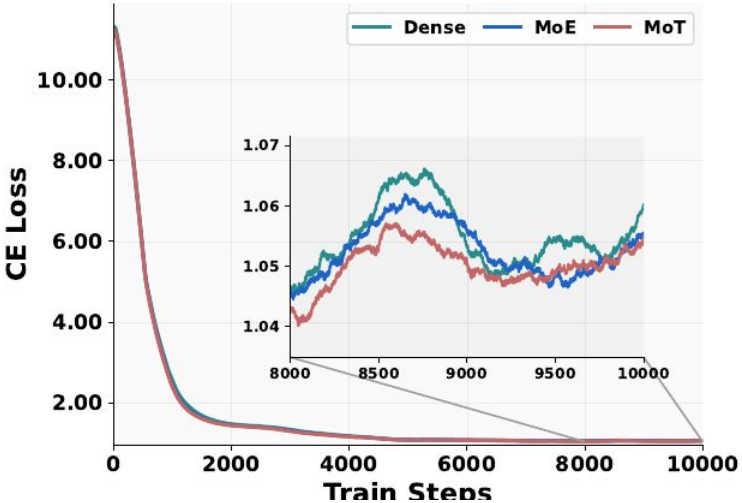
Architecture

- Qwen2.5 copied into two experts
- Separate weights, shared attention
- Generated images appear twice in training: noisy (for the loss) and clean (what later tokens see)

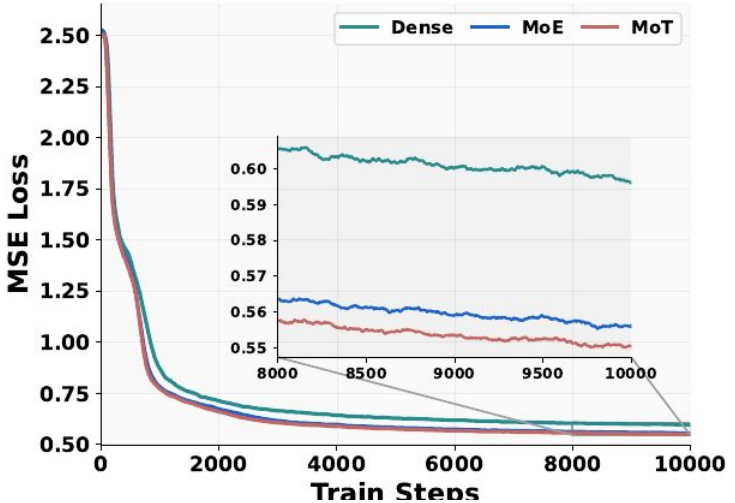


Why Two Experts?

Dense vs MoE vs MoT, same FLOPs, 1.5B model, only the architecture changes



understanding (CE)



generation (MSE)

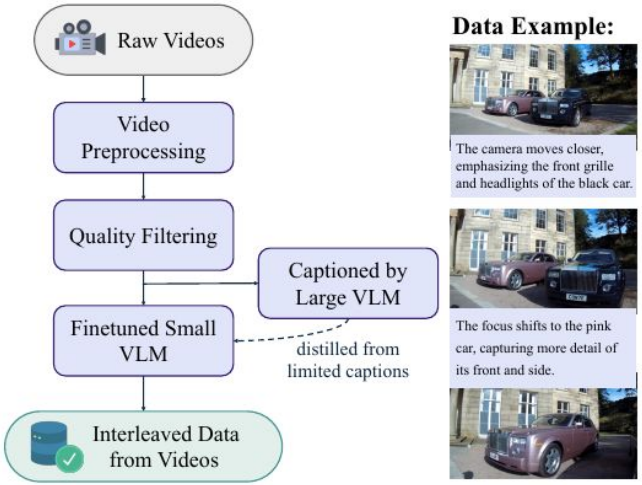
- MoE: copy only the FFN. MoT: copy everything
- Separate generation weights help a lot, full MoT a bit more

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

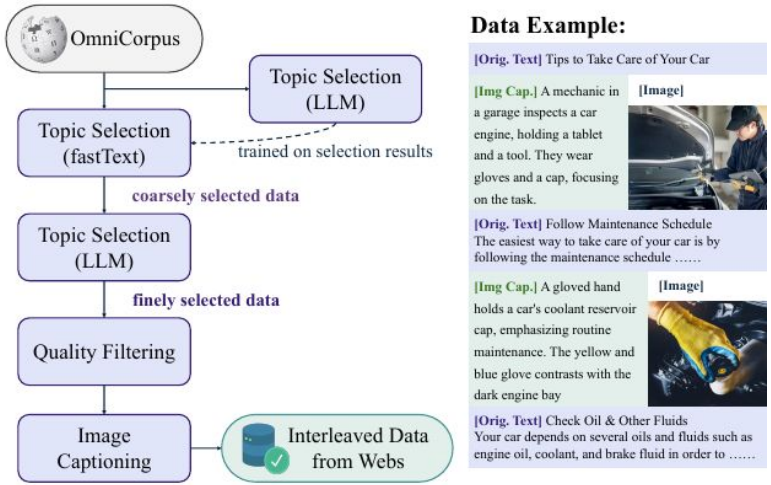
Data

- Video → frames + short “what changed” captions (45M)
- Web → caption written before each image (20M docs)
- Plus 500k reasoning examples (<think> before generating)

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



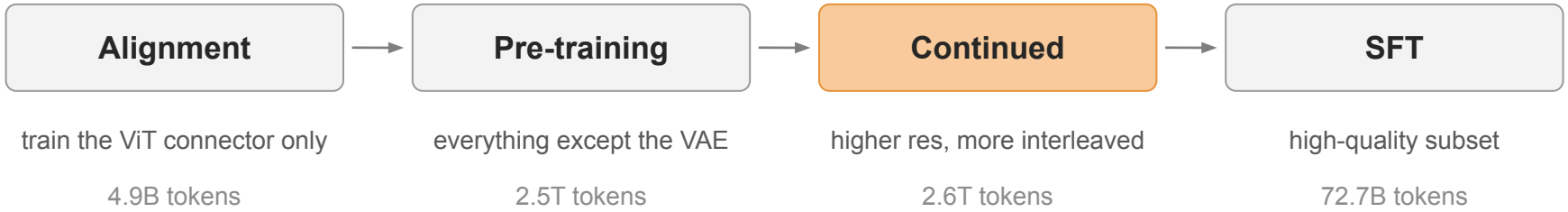
(a) Data pipeline for interleaved data from videos.



(b) Data pipeline for interleaved data from webs.

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Training Recipe



Two things they learned (on a 1.5B model):

- Sample generation data way more often (~4 : 1)
- Big learning rate helps generation, hurts understanding → weight the losses (CE : MSE = 0.25 : 1)

Results

How They Evaluate

Task	Benchmarks	Scored by
Understanding	MME, MMBench, MMMU, MM-Vet, MathVista, MMVP	standard metrics
Text-to-image	GenEval (composition), WISE (world knowledge)	standard metrics
Editing	GEdit-Bench (real user edit requests)	GPT-4.1 judge
Reasoning edits	IntelligentBench (authors' own, 350 examples)	GPT-4o judge, 0–2 each

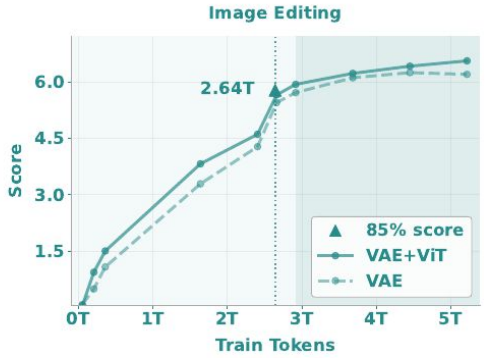
- Baselines: open unified (Janus-Pro, MetaQuery-XL), specialists (Qwen2.5-VL, FLUX, Step1X-Edit), private (GPT-4o, Gemini 2.0)
- Also re-evaluate old checkpoints to see when each ability shows up

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

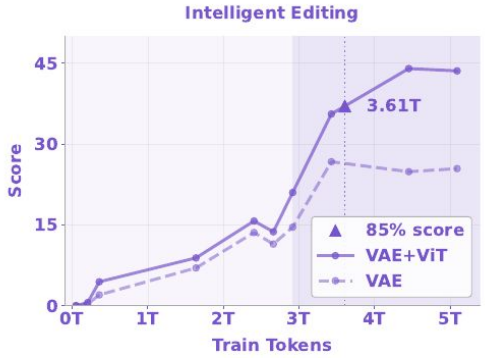
Emerging Properties

- Tokens to reach 85% of final score:
  - Understanding: 0.18T
  - Generation: 0.68T
  - Editing: 2.64T
  - Intelligent editing: 3.61T
- No ViT tokens → 16% drop on intelligent editing

dashed = VAE only (no ViT), shaded = continued-training stage



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



(d) IntelligentBench Score on Intelligent Editing task.

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Emergence, Visually

Same question, checkpoints at 1.5T → 4.5T tokens



Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Benchmarks

	BAGEL	Open unified	Specialist	Private
MMMU	**55.3**	58.6 MetaQuery-XL†	58.6 Qwen2.5-VL	–
MM-Vet	**67.2**	66.6 MetaQuery-XL†	67.1 Qwen2.5-VL	–
GenEval	**0.88***	0.80 Janus-Pro	0.82 FLUX.1-dev	–
GEdit-EN	**6.52**	–	6.70 Step1X-Edit	7.53 GPT-4o
IntelligentBench	**44.9**	–	14.9 Step1X-Edit	78.9 GPT-4o‡ 57.6 Gemini 2.0

† frozen Qwen2.5-VL backbone * with LLM prompt rewriter (0.82 without) ‡ only the 318/350 questions it answered

Roughly ties the specialists, still well behind private models on reasoning edits

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Thinking Helps

- Let BAGEL write a <think> before it draws
- WISE (world-knowledge T2I): 0.52 → 0.70
- IntelligentBench: 44.9 → 55.3

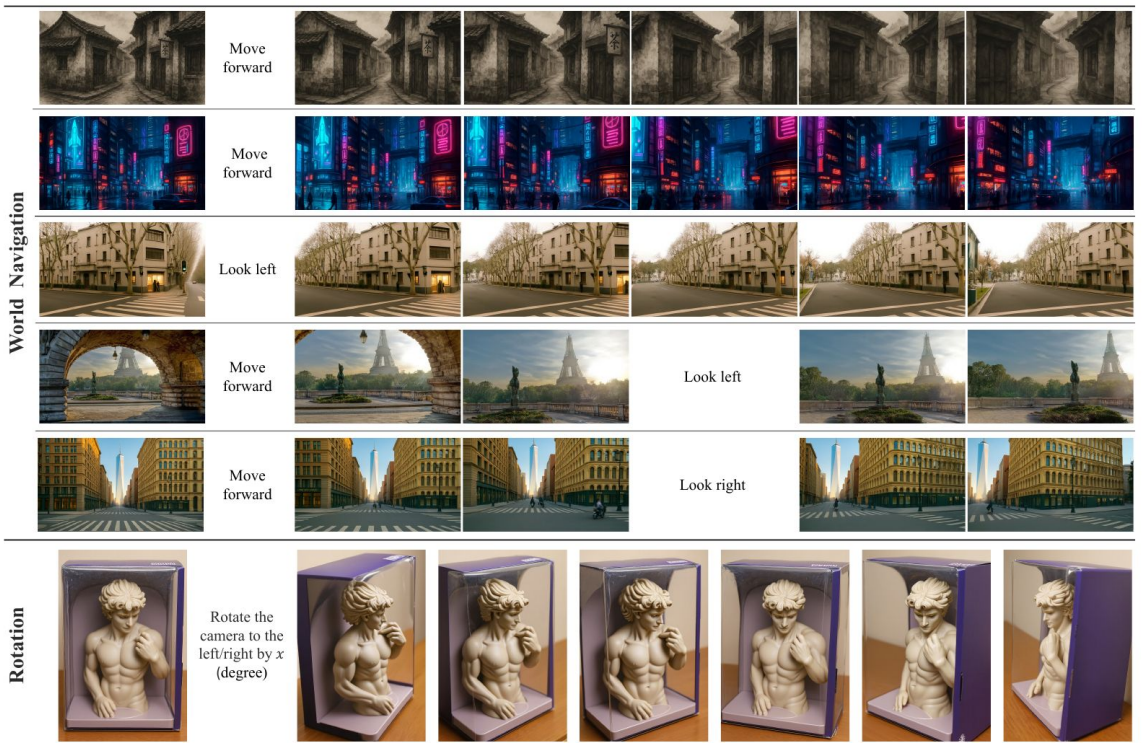
For reference:
GPT-4o 0.80 WISE
Gemini 2.0 57.6 IntelligentBench

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>**	

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

World Modeling

- Fine-tuned with more video + navigation data
- “Move forward”, “look left”, rotate the object
- Trained on real streets, works on paintings, cartoons, games
- (qualitative only)



Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Strengths

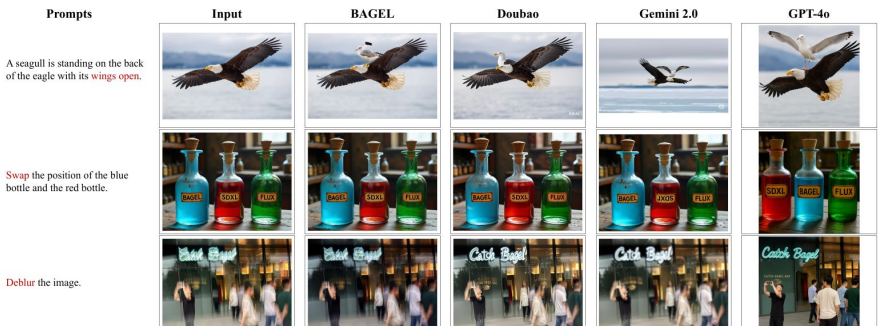
- Actually open: weights, code, data recipe
- Useful training lessons (expert split, data ratios, learning rates)
- Tracks abilities over training, not just the final checkpoint
- One model: understand, generate, edit, navigate
- No bottleneck design (their argument, not ablated)



Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Weaknesses

- Is it really “emergence”?
  - Jump at ~3T coincides with more interleaved data
- Own benchmark, graded by GPT-4o
- Expert ablation only at 1.5B
- Still behind GPT-4o and Gemini 2.0 on reasoning edits
- Hard to reproduce (~5T tokens, data not released)



Failure cases: edits like swapping objects are still hard

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Open Questions

- Why do the two experts need full separate weights? What do the experts learn that shared attention can't?
- Why keep both ViT and VAE tokens? Removing ViT costs 16% on intelligent editing; is that information the VAE path simply can't recover?
- Which paradigm is more likely to dominate future unified multimodal models: a fully shared Transformer or specialized experts with shared attention, as in BAGEL?
- Is “emerging” the right word here, or is it just that harder tasks take longer to learn?
- How much of the gain comes from the interleaved data versus just training longer on more stuff?

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Open Questions

- Why do the experts need fully separate weights?
- Why keep both ViT and VAE tokens?
- Fully shared Transformer, or experts + shared attention?
- “Emerging”, or just harder tasks taking longer?
- Interleaved data, or just more training?

Deng, Chaorui, et al. "Emerging Properties in Unified Multimodal Pretraining." arXiv:2505.14683 (2025)

Questions?