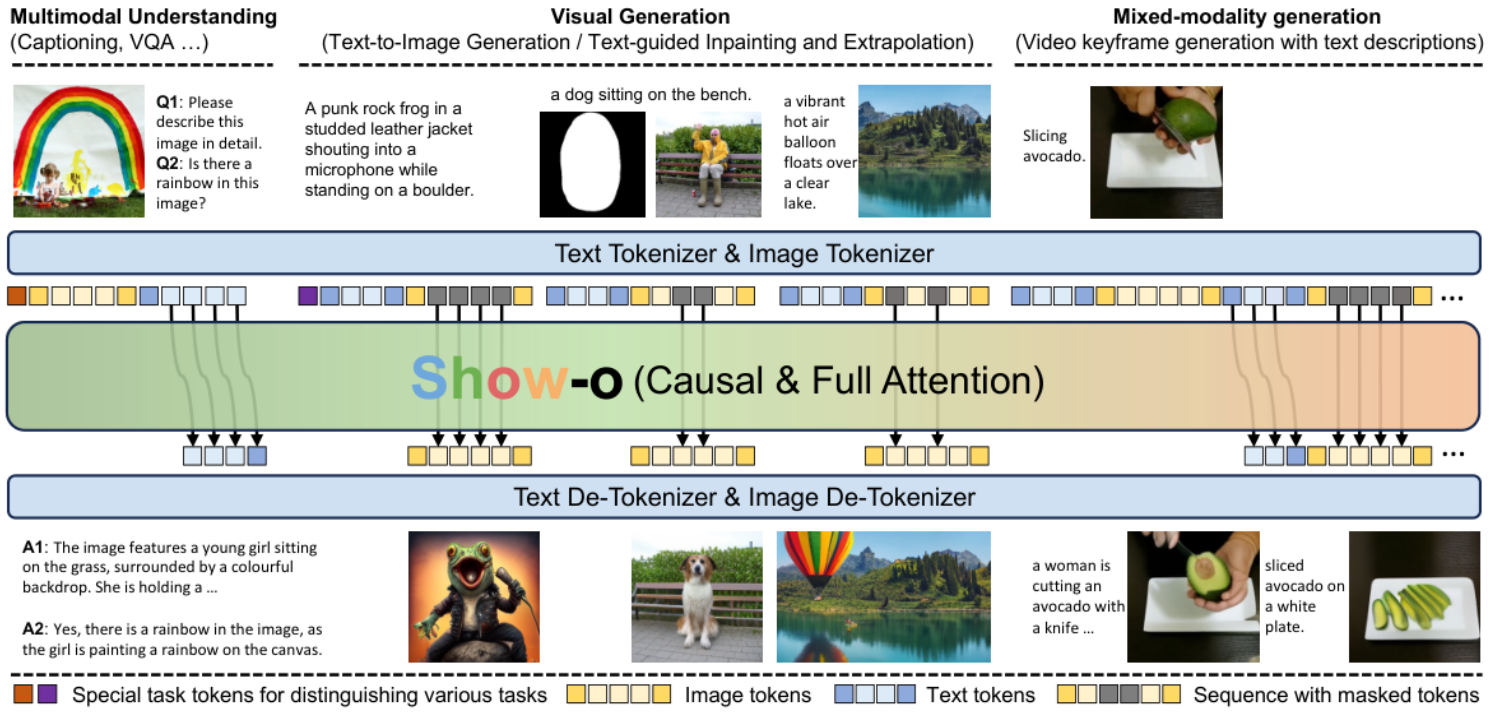
Show-o

One Single Transformer to Unify Multimodal Understanding and Generation



Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang et al.

Presenter: Yidong Huang

Why unify understanding and generation?

One image can be an input to understand — or an output to create.

Image → text

“What is in this image?”
Predict answer tokens autoregressively.

Text → image

“Generate an image of a red bicycle.”
Fill visual tokens by masked refinement.

Show-o: a shared transformer, with the decoding rule matched to the modality.

Problem: understanding and generation are usually separate

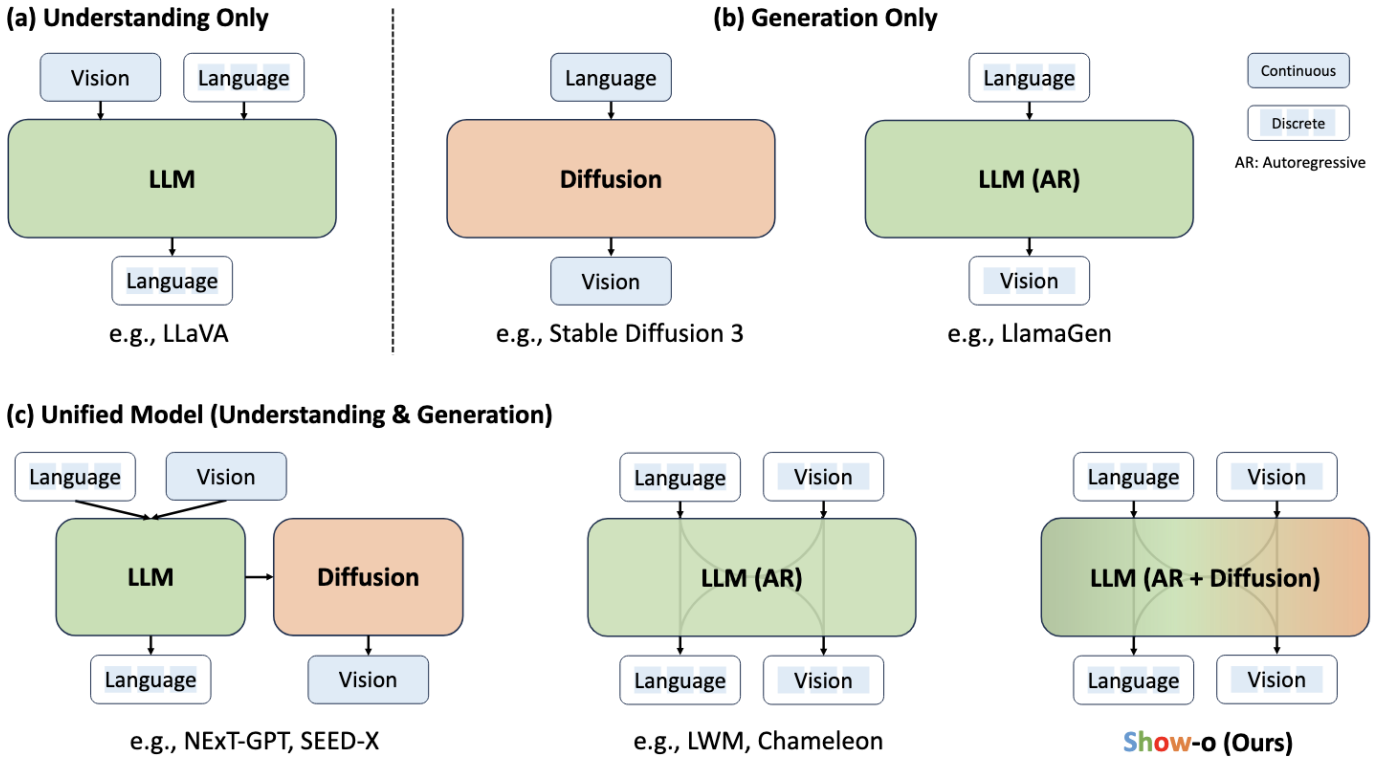
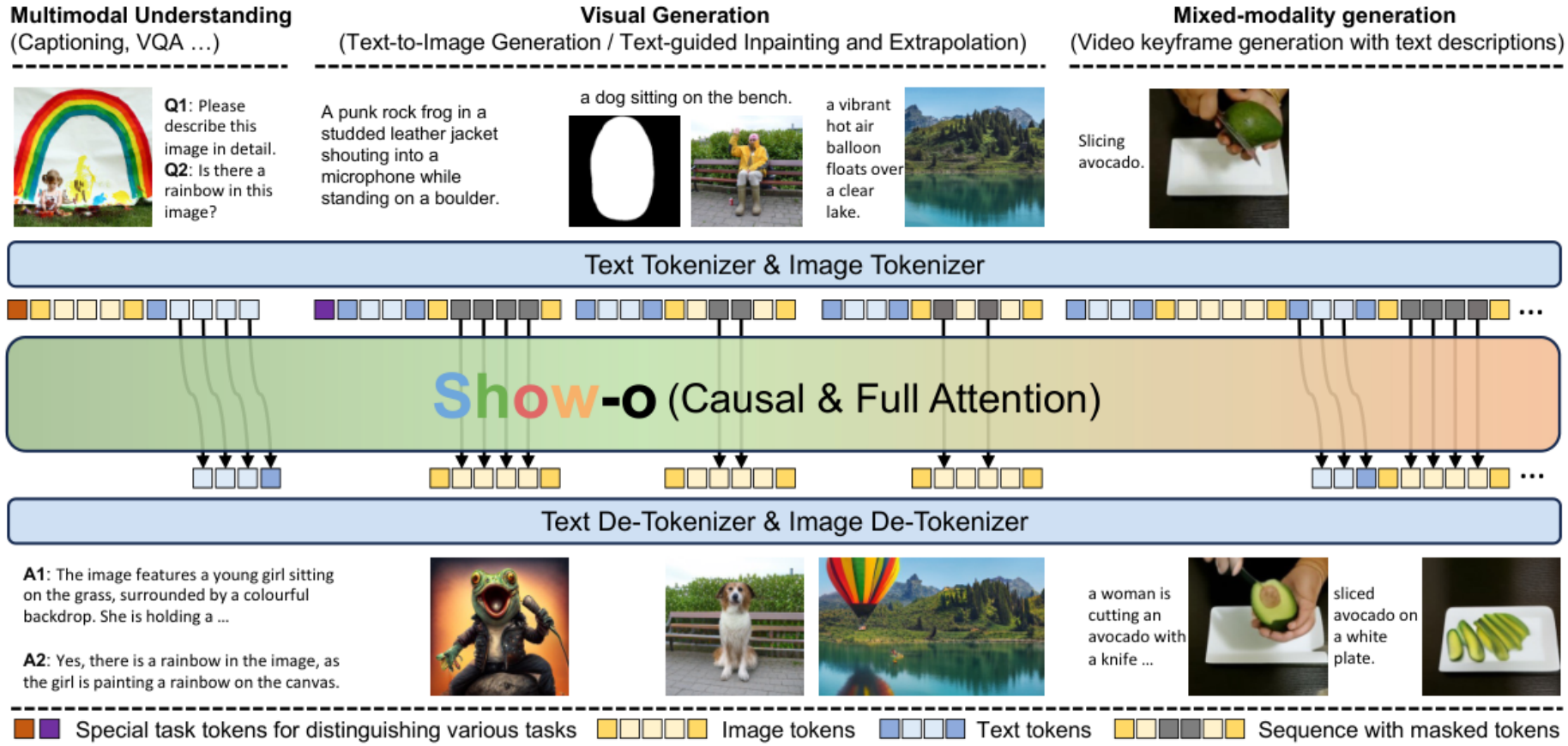


Figure 1: Characteristics comparison among understanding only, generation only, and unified (understanding & generation) models. "Vision" and "Language" indicate the representations from specific input modalities. In this context, "Diffusion" represents both continuous and discrete diffusion.

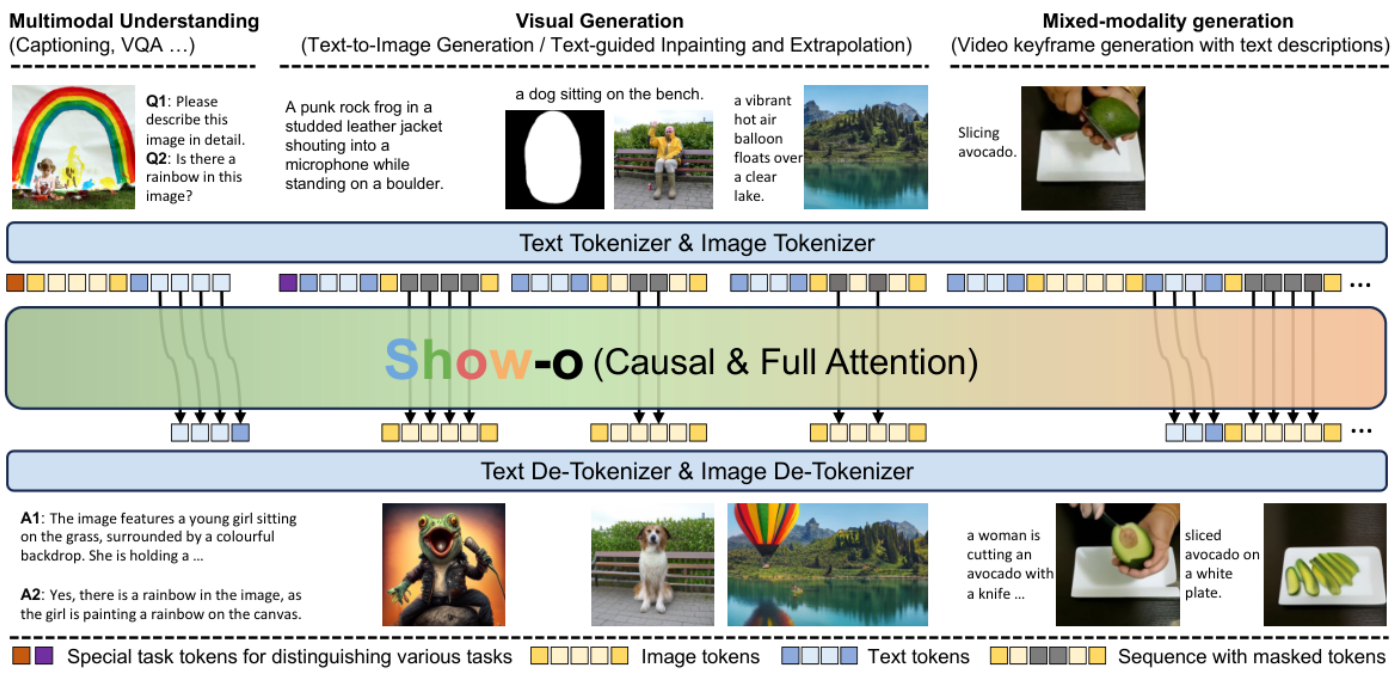
- Understanding: image-conditioned text prediction.
- Generation: text-conditioned image-token refinement.

Architecture: the same transformer for all tasks



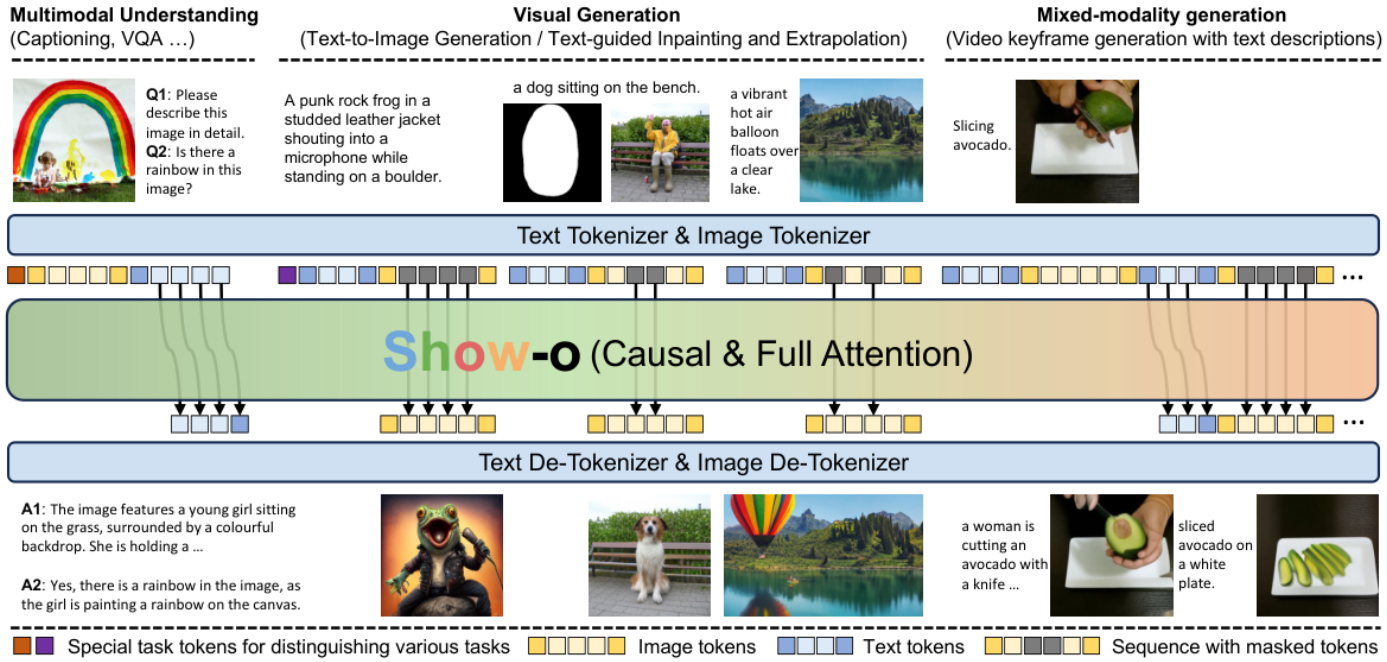
Autoregressive text tokens + masked diffusion-style image tokens.

A common discrete token space



- Text: standard Phi-1.5 vocabulary.
- Images: MAGVIT-v2-style lookup-free quantizer.
- 8,192 image codebook entries.
- 256 × 256 image → 16 × 16 = 256 image tokens.

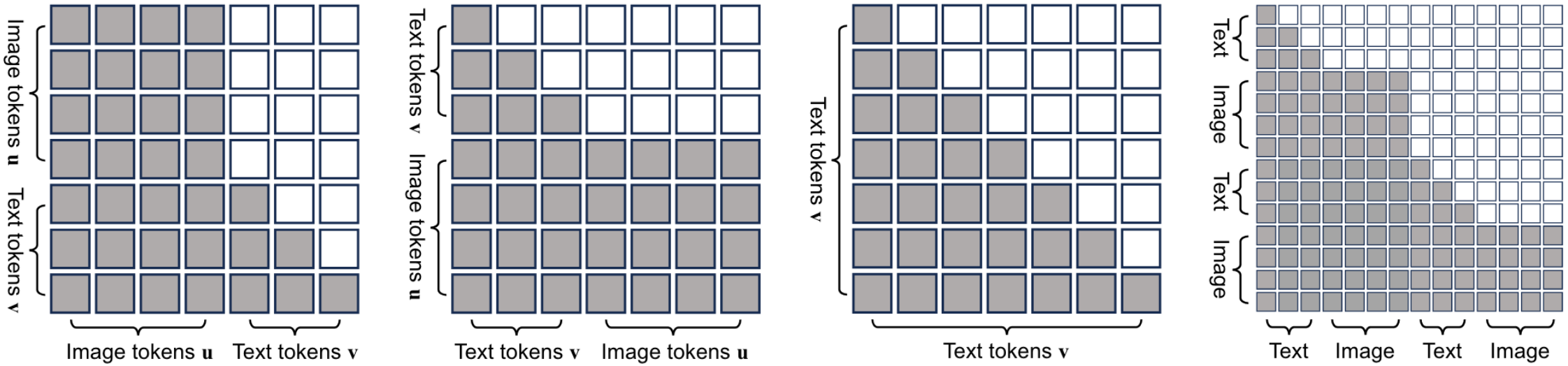
Unified prompting chooses the behavior



Special task markers

- [MMU]: image understanding, then autoregressive answer.
- [T2I]: text prompt, then masked image-token prediction.

Omni-attention: different attention masks by modality



- Text positions: causal attention for next-token prediction.
- Image positions: bidirectional attention for masked-token recovery.

Training: two prediction objectives, one transformer

1 Understanding (image → text)

Image context IMG Q A_1 A_2 A_3

Predict each answer token from its causal prefix (NTP)

2 Generation (text → image)

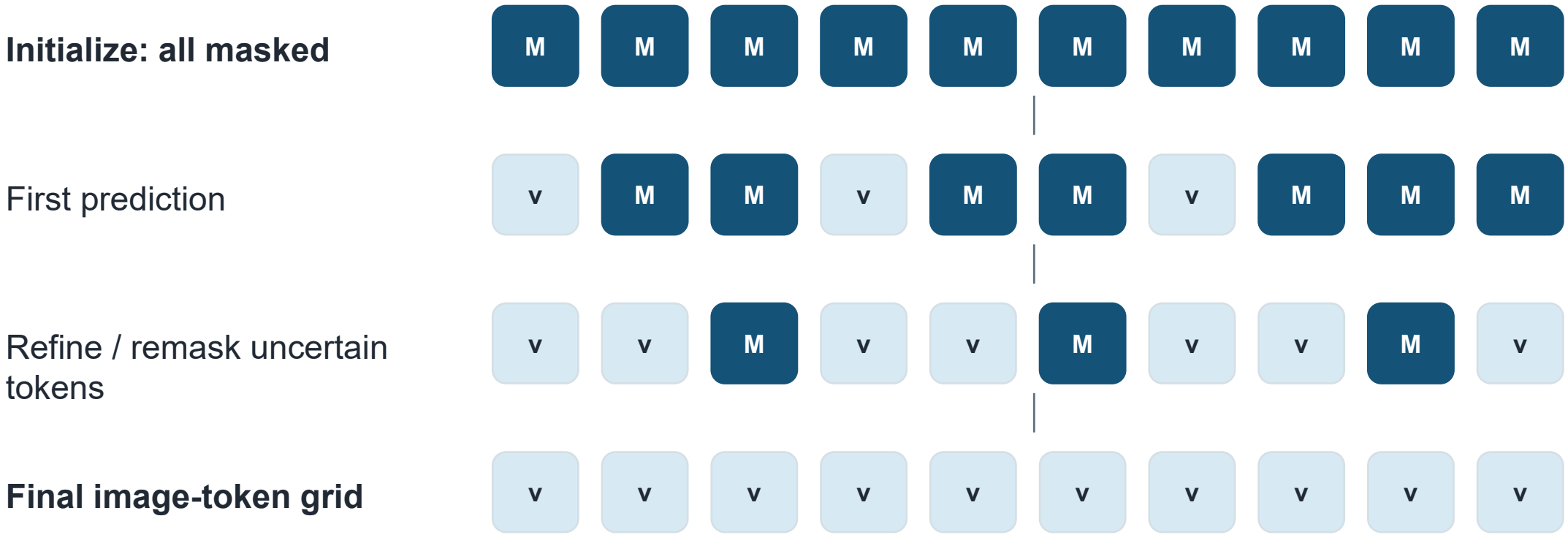
Text prompt TXT v_1 M v_3 M M M v_7 M

Randomly mask image-token positions; predict missing tokens (MTP)

Joint loss: $L = L_{MTP} + \alpha L_{NTP}$

Inference: image tokens are refined in several passes

Condition on the text prompt; predict masked image tokens jointly.



Unlike left-to-right image autoregression, multiple visual tokens can be updated each pass.

Experimental protocol

- Base backbone: pretrained Phi-1.5 (~1.3B parameters).

- Pretraining resources: 2.0B image–text pair scales;(Imagenet, DataComp, COYO700M)
- Finetuning: 35M fine-tuning examples (CC12M + SA1B + LAION)

- Understanding: POPE, MME, Flickr30k, VQAv2, GQA, MMMU.

- Generation: MSCOCO zero-shot FID and GenEval compositional alignment.

Important control: the paper trains an understanding-only Phi-1.5 + LLaVA baseline.

Multimodal understanding: competitive, but not uniformly best

Table 1: Evaluation on multimodal understanding benchmarks. Show-o is currently built upon Phi-1.5 (Li et al., 2023) and thus we implement LLaVA-v1.5-Phi-1.5 as our apple-to-apple baseline. Und. and Gen. denote "understanding" and "generation", respectively. ‡ denotes the improved Show-o that employs CLIP-ViT continuous representations. We highlight the model size of Show-o and LLaVA baseline in green, and we use blue to highlight the larger model size than ours.

Type	Model	# Params	POPE↑	MME↑	Flickr30k↑	VQAv2$_{(test)}$↑	GQA↑	MMMU↑
Und. Only	LLaVA-v1.5	7B	85.9	1510.7	-	78.5	62.0	35.4
	InstructBLIP	13B	78.9	1212.8	-	-	49.5	-
	Qwen-VL-Chat	7B	-	1487.5	-	78.2	57.5	-
	mPLUG-Owl2	7B	85.8	1450.2	-	79.4	56.1	-
	LLaVA-v1.5-Phi-1.5	1.3B	84.1	1128.0	69.6	75.3	56.5	30.7
Und. and Gen.	Gemini-Nano-1	1.8B	-	-	-	62.7	-	26.3
	CoDI	-	-	-	12.8	-	-	-
	Emu	13B	-	-	77.4	57.2	-	-
	NExT-GPT	13B	-	-	84.5	66.7	-	-
	SEED-X	17B	84.2	1435.7	52.3	-	47.9	35.6
	Chameleon	34B	-	-	74.7	66.0		
	Show-o (Ours)	1.3B	73.8	948.4	36.2	59.3	48.7	25.1
	Show-o‡ (Ours)	1.3B	84.5	1182.7	64.3	73.9	57.5	27.4

VQAv2 (test): 69.4 (discrete Show-o) | 74.7 (CLIP-ViT variant) | 75.3 (LLaVA-Phi).

Text-to-image generation: composition benchmarks

Table 3: Evaluation of text-to-image generation on the GenEval (Ghosh et al., 2023) benchmark. Und. and Gen. denote "understanding" and "generation", respectively. We highlight the model size of Show-o in green, and we use blue to highlight the larger model size than ours. Obj.: Object. Attri.: Attribute.

Type	Method	# Params	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	**Overall↑**
Gen. Only	LlamaGen	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM	1.4B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5	0.9B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-alpha	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1	0.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2	6.5B	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	SDXL	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	SD3 (d=24)	2B	0.98	0.74	0.63	0.67	0.34	0.36	0.62
Und. and Gen.	CoDI	-	0.89	0.16	0.16	0.65	0.02	0.01	0.31
	LWM	7B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o (Ours)	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53

GenEval overall: Show-o 0.68 | Show-o (CLIP variant) 0.69 | SD3 (2B) 0.62.

Image quality and visual examples



- Text-to-image outputs can capture styles, objects and scenes.
- MSCOCO zero-shot FID-30K: 9.24.
- A qualitative figure cannot replace human preference or robustness tests.

Xie et al., ICLR 2025 • Fig. 7

Inpainting emerges without task-specific fine-tuning



- Mask just the target visual-token region and predict replacement codes.
- Prompt-conditioned inpainting and extrapolation use the same model.

Xie et al., ICLR 2025 • Fig. 8 / project examples

Multimodal understanding: unusual visual questions



User: Do you think the image is unusual or not?

Show-o:
The image is unusual because it depicts a living room with a couch and a table in the middle of the ocean, surrounded by a large body of water. This is not a typical setting for a living room, as living rooms are usually indoors and designed for relaxation and entertainment. The presence of the couch and table in the ocean adds a unique and unexpected element to the scene, making it stand out from typical living room setups.

Chameleon:
The image is unusual because of the transparent and colorful fish and coral in the water. The image is also unusual because of the clear blue sky with white clouds. The unusual objects in the image are the fish and coral in the water. The unusual objects in the image are the fish and coral in the water. The unusual objects in the image are the fish and coral in the water. The unusual objects in the image are the fish and coral in the water.

SEED-X:
Yes, the image is unusual. It features a living room set, including a couch and two chairs, placed in the middle of the ocean. The room is surrounded by water, with coral reefs and fish visible in the background. This is a surreal and unexpected scene, as living rooms are typically found on land, not in the middle of the ocean. The presence of the couch and chairs in such an environment adds to the peculiarity of the image.

- Original VQA comparison with Chameleon.
- Show-o can explain unusual image semantics.
- Case studies are illustrative—not controlled benchmarks.

Strengths and limitations

Strengths

- One transformer, mixed-modality prompting and shared parameters.
- Fast non-autoregressive image-token decoding.
- Natural inpainting and mixed-modality applications.

Limitations

- Discrete tokenization introduces a compression / representation bottleneck.
- Default discrete understanding trails the CLIP-ViT input variant.
- Results are at moderate visual resolution and relatively small scale.

Follow-up: Show-o2 scales beyond discrete visual tokens

Show-o (2025)

- Discrete image tokenizer
- Masked-token image generation
- Primarily image + text tasks

Show-o2 (2025)

- Continuous 3D causal VAE features
- Flow matching for image / video generation
- Autoregressive language head + flow head
- 1.5B and 7B configurations

Research direction: choose the right objective and representation for each modality.

Continuous vs. Discrete Image Tokens

Is the quality gap mainly limited training data or information loss from quantization?

Answer: Both matter. Quantization can impose a fidelity ceiling; data / compute govern how well the generator models those tokens.

ISOLATE THE TOKENIZER

TokenBridge (Wang et al., ICCV 2025), Table 1(a)

Representation	Levels B	rFID ↓
Continuous VAE	—	**1.11**
Discrete (post-quantized)	8	**3.69**
Discrete (post-quantized)	16	**1.33**
Discrete (post-quantized)	32	**1.12**
Discrete (post-quantized)	**64**	**1.11**

ImageNet train split; same VAE and decoder, varying quantization levels.

WHAT THIS MEANS

1. **Quantization loss is real**

   Coarse discrete codes lose details before the generator is even trained; more generator data cannot undo that fixed tokenizer loss.

2. **But discreteness is not the culprit itself**

   Increasing granularity from 8 → 64 levels closes the reconstruction gap (rFID 3.69 → 1.11).

3. **Data, compute, and modeling still matter**

   For small generators, more compressed latents may be easier to model—even when their reconstruction is worse (Ramanujan et al., NeurIPS 2025).

Diagnostic: compare reconstruction quality (tokenizer) separately from generation quality (model + training).

Sources: Wang et al., ICCV 2025 (arXiv:2503.16430), Table 1(a); Ramanujan et al., NeurIPS 2025 (arXiv:2412.16326).

Discussion: what should unified models share?

01 Do images need discrete tokens, or should generation use continuous flow matching?

02 What should be shared: the transformer backbone, attention pattern (kv cache), or generation head?

03 When does joint training help—or cause interference between understanding and generation? (Towards Physics of Multimodal Pretraining)

Paper: arxiv.org/abs/2408.12528

Author project: showlab.github.io/Show-o