Harmonizing Visual Representations for Unified Multimodal Understanding and Generation

Size Wu[1] Wenwei Zhang[2] Lumin Xu[3] Sheng Jin[4]
Zhonghua Wu[5] Qingyi Tao[5] Wentao Liu[4] Wei Li[1] Chen Change Loy[1]

[1]S-Lab, Nanyang Technological University [2]Shanghai AI Laboratory
[3]The Chinese University of Hong Kong [4]Sensetime Research and Tetras.AI [5]Sensetime Research

Outline

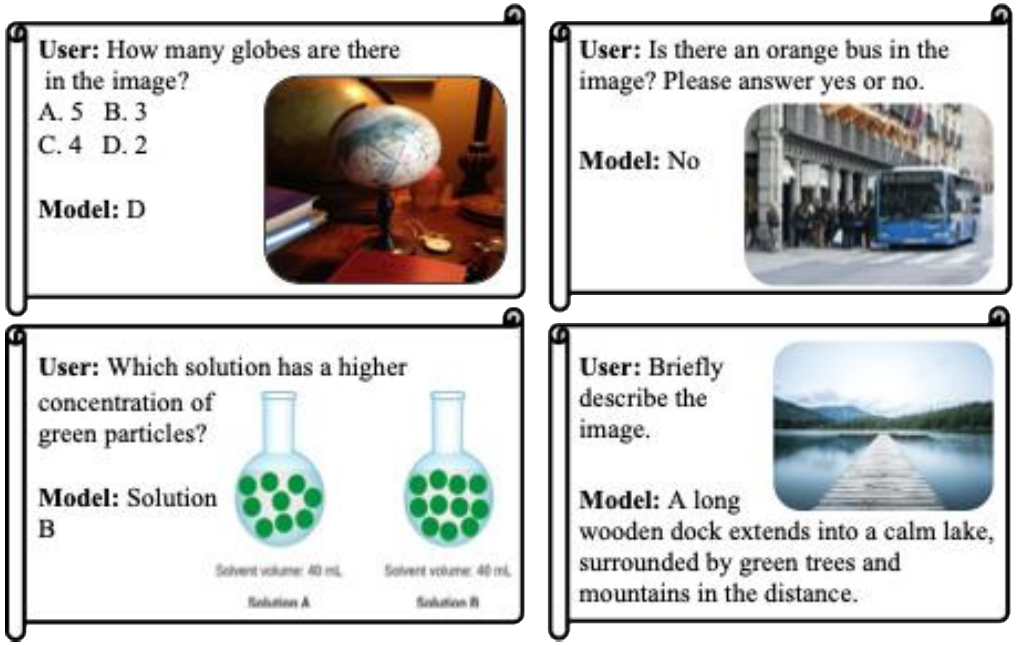
- Background
- Introduction
- Method
- Results

Outline

- Background
- Introduction
- Method
- Results

Background

- Towards Unified Multimodal Models



User: How many globes are there in the image?
A. 5 B. 3
C. 4 D. 2

Model: D

User: Is there an orange bus in the image? Please answer yes or no.

Model: No

User: Which solution has a higher concentration of green particles?

Model: Solution B

User: Briefly describe the image.

Model: A long wooden dock extends into a calm lake, surrounded by green trees and mountains in the distance.

Solvent volume: 40 mL
Solution A
Solvent volume: 40 mL
Solution B

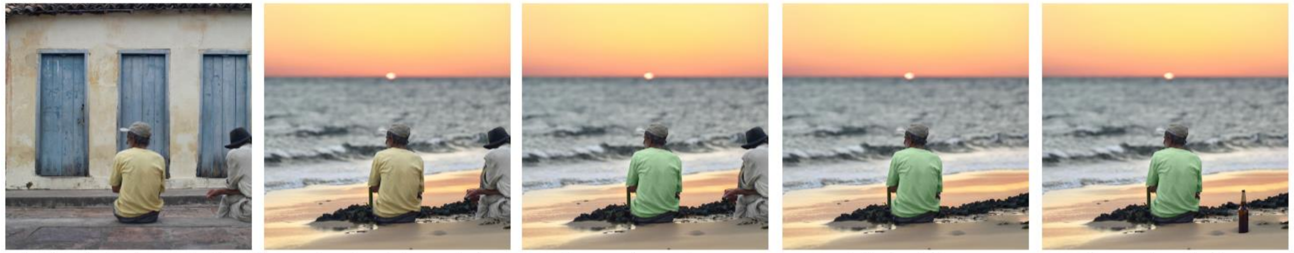
(a) Image understanding

(b) Image generation

Background

- Towards Unified Multimodal Models

Multi-turn Editing



Replace the background with a beach under sunset.

Change the color of the men's clothes at the middle bottom to light green.

Remove the man in the bottom right.

Add a bottle in the bottom right.

Change the style of the image to pen art.

Story Telling

Here is the beginning of a story about George. Please continue it: The man led George into a nearby building.

Feeling tired from their adventure, George climbed onto the bed, snuggling under the blanket. The room was peaceful, and George was quickly drifting off to sleep.

George followed the man into a colorful hallway. The man was gesturing towards a closed door, and George was curious about what was behind it.

George found a cream-colored cat with a collar. He gently stroked the cat's head, glad to have found a new friend.

George decided to take the cat for a walk. He held the cat by the collar, looking quite pleased with himself.

Multimodal Assistant

Step 1: Crack the eggs into a bowl and whisk them together with a fork or a whisk.

Step 2: Heat a non-stick pan or skillet over medium heat.

Step 3: Once the oil is hot, pour the eggs into the pan and use a spatula to gently scramble them.

Step 4: Once the eggs are cooked to your desired doneness, season with salt and pepper to taste.

Step 5: Once the eggs are cooked to your liking, transfer them to a serving plate and enjoy!

Complex conversational scenarios with understanding and generation.

Background

- Towards Unified Multimodal Models

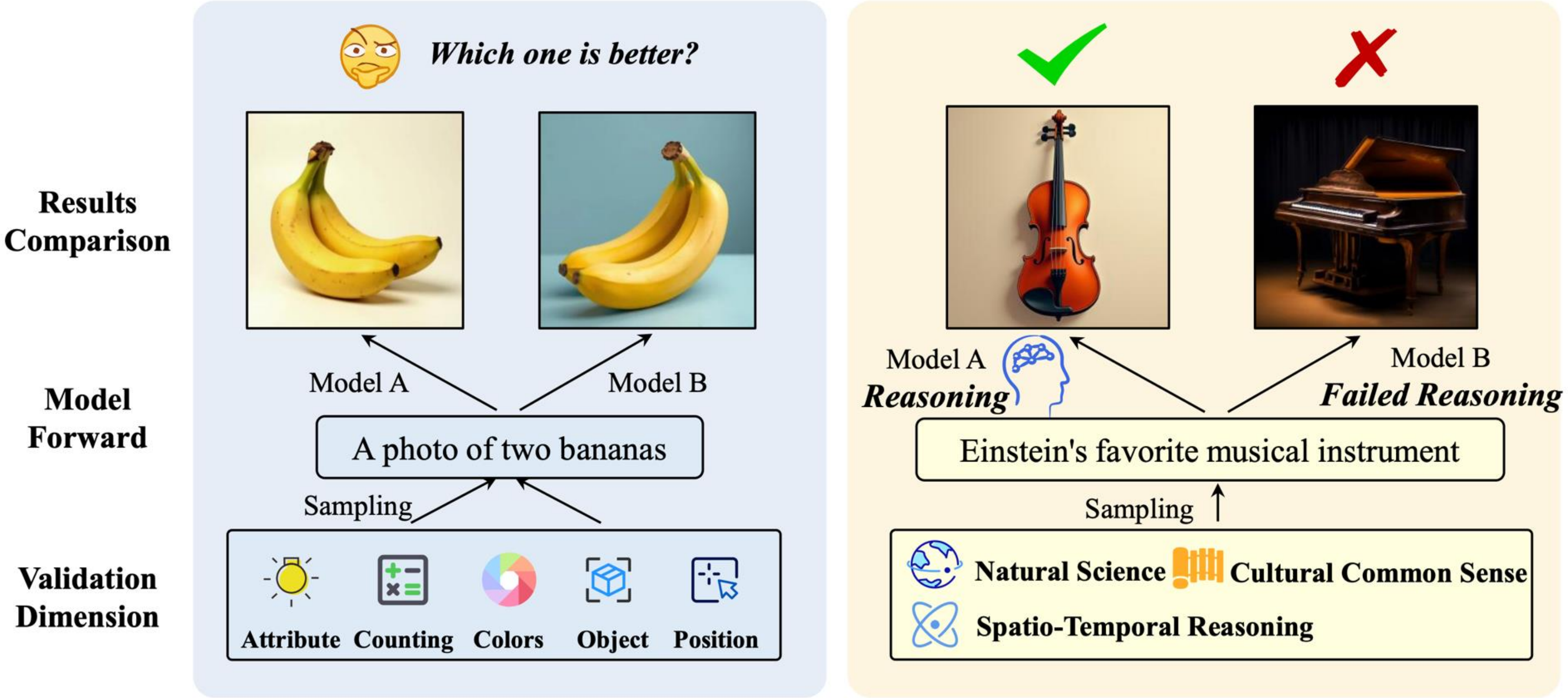


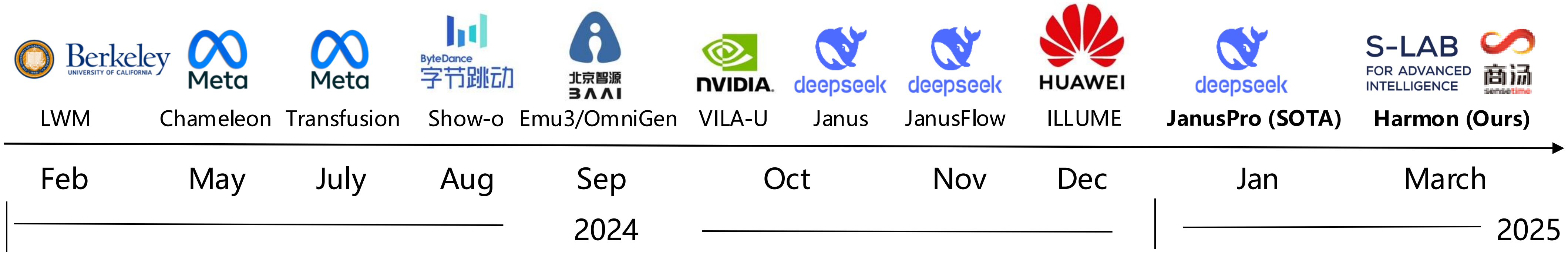
Image generation with world knowledge. Examples are taken from WISE [1].

[1] WISE: A World Knowledge-Informed Semantic Evaluation for Text-to-Image Generation, Niu et.al., Arxiv 2025

Background

- Towards Unified Multimodal Models

"What I cannot create, I do not understand" ---Richard Feynman



LWM Chameleon Transfusion Show-o Emu3/OmniGen VILA-U Janus JanusFlow ILLUME **JanusPro (SOTA)** **Harmon (Ours)**

Feb May July Aug Sep Oct Nov Dec Jan March

2024 2025

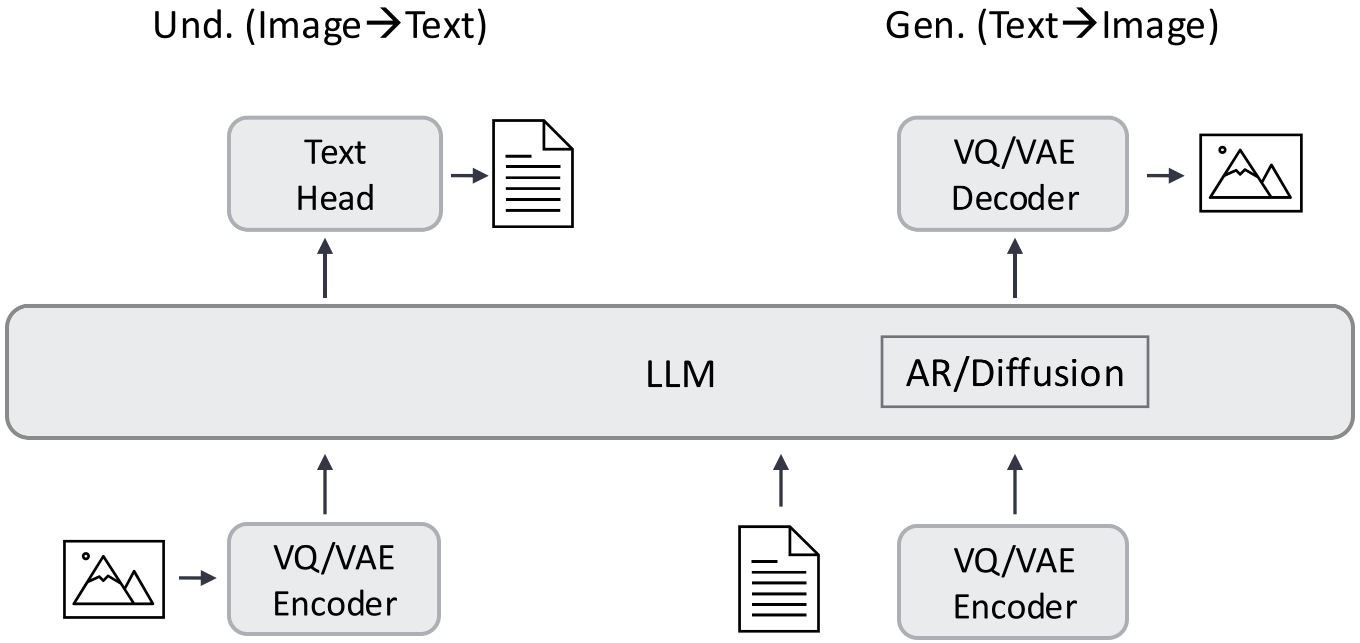
The emergence of unified multimodal models since 2024.

Outline

- Background
- Introduction
- Method
- Results

Introduction

- Existing paradigms

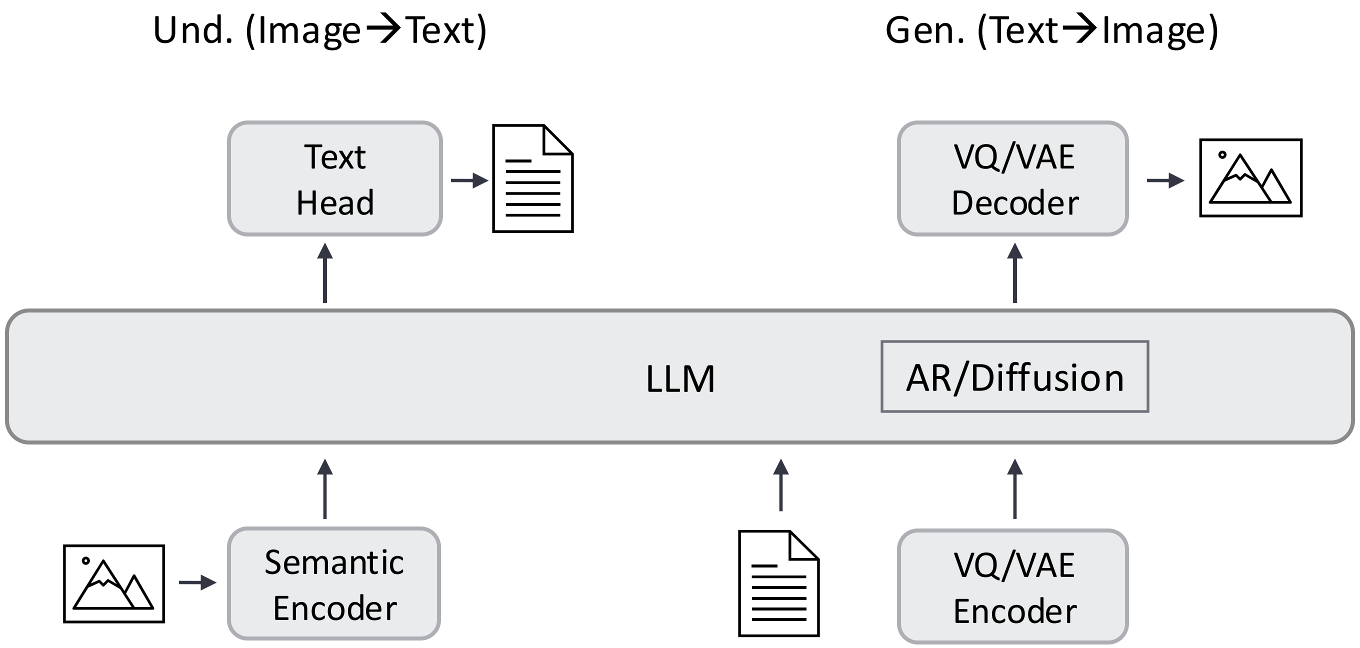


Representing images as VQ tokens or VAE latents [2, 3, 4].

[2] Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, Zhou et.al., ICLR 2025
[3] Show-o: One Single Transformer to Unify Multimodal Understanding and Generation, Xie et.al., ICLR 2025
[4] VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation, Wu et.al., ICLR 2025

Introduction

- Existing paradigms

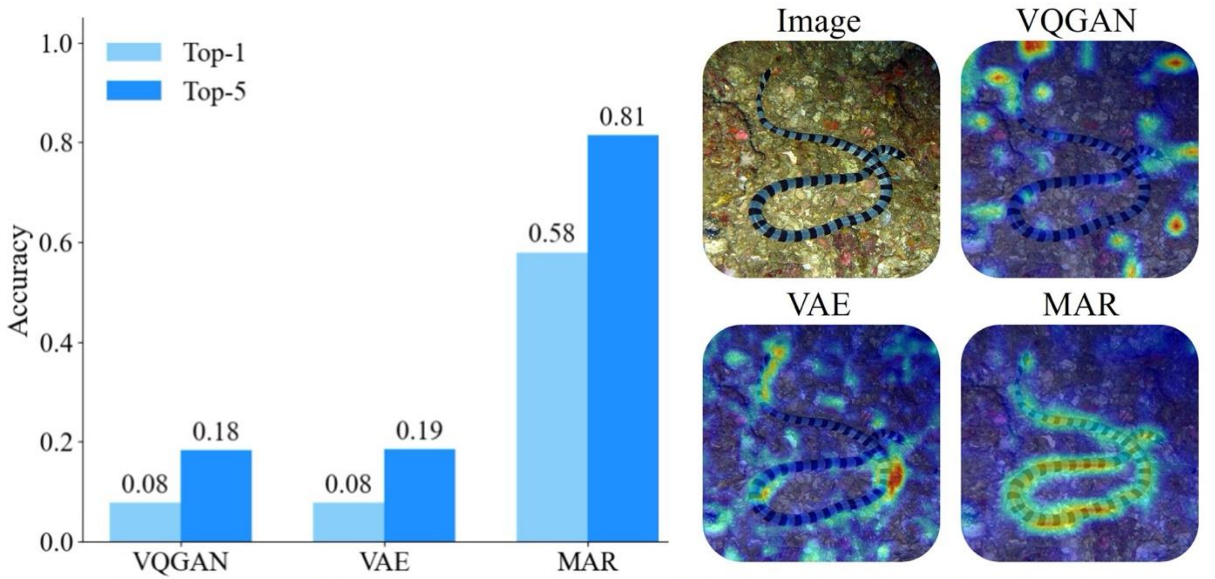


Representing images as semantic embeddings for understanding and VQ tokens/VAE latents for generation [5, 6, 7].

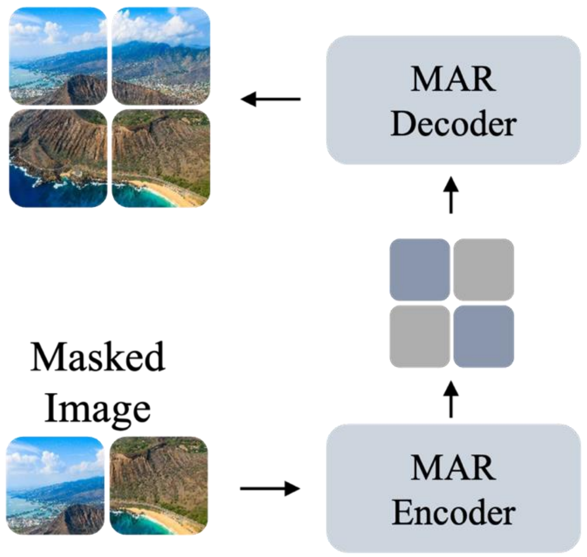
[5] Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation, Wu et.al., CVPR 2025
[6] JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation, Ma et.al., CVPR 2025
[7] Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, Chen et.al., Arxiv 2025

Introduction

- Our observation: masked image modelling harmonizes und. & gen.



(a) Linear probing on ImageNet (b) Visualization of activation maps



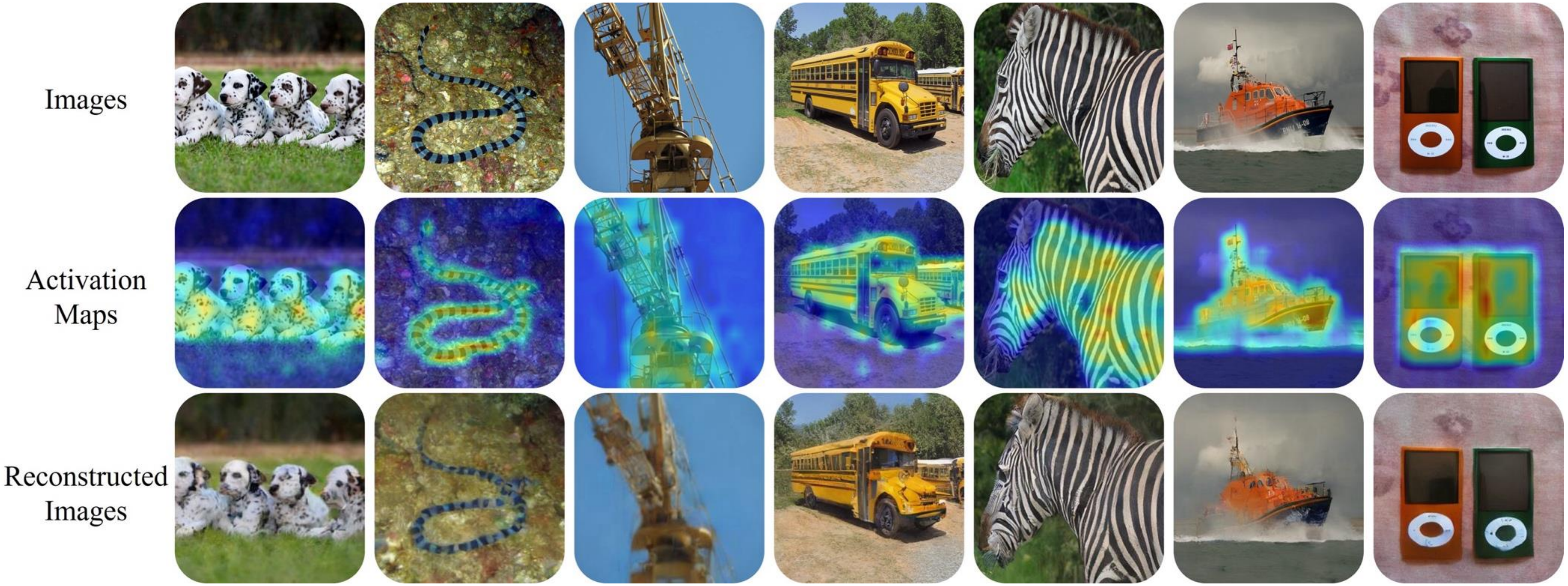
MAR: masked autoregression [8].

---

[8] Autoregressive Image Generation without Vector Quantization, Li et.al., Neurips 2024

Introduction

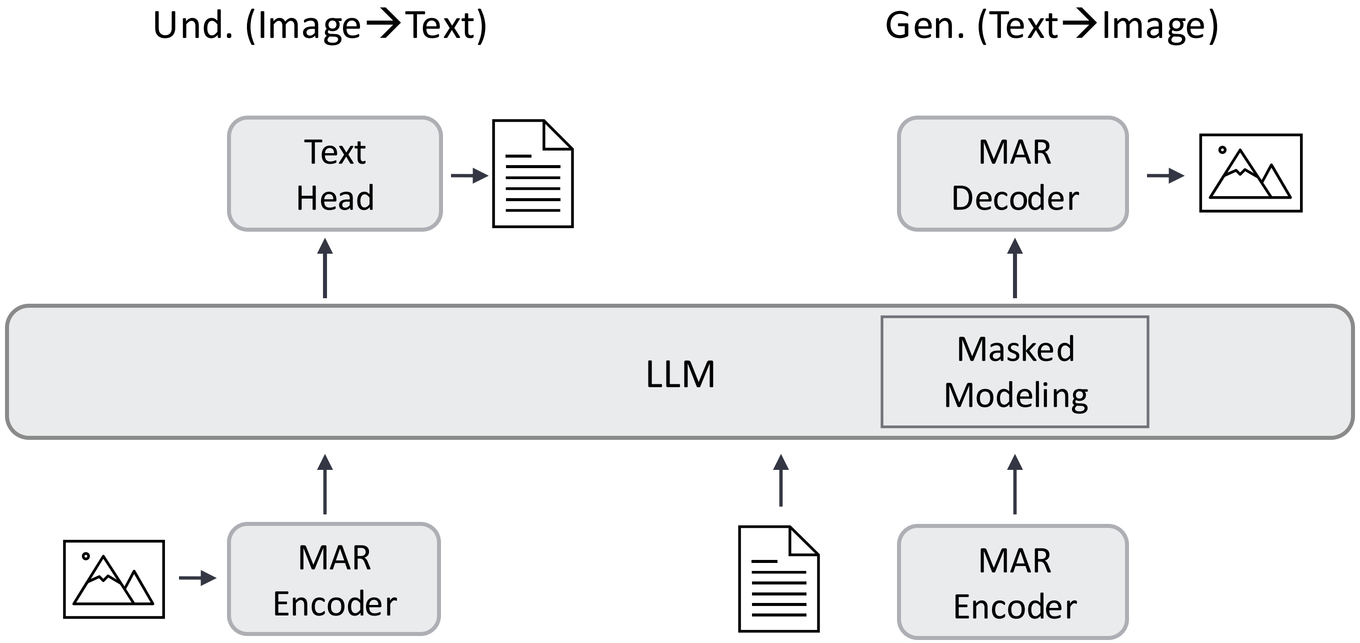
- Our observation: masked image modelling harmonizes und. & gen.



Visualization of MAR feature maps.

Introduction

- Our observation: masked image modelling harmonizes und. & gen.



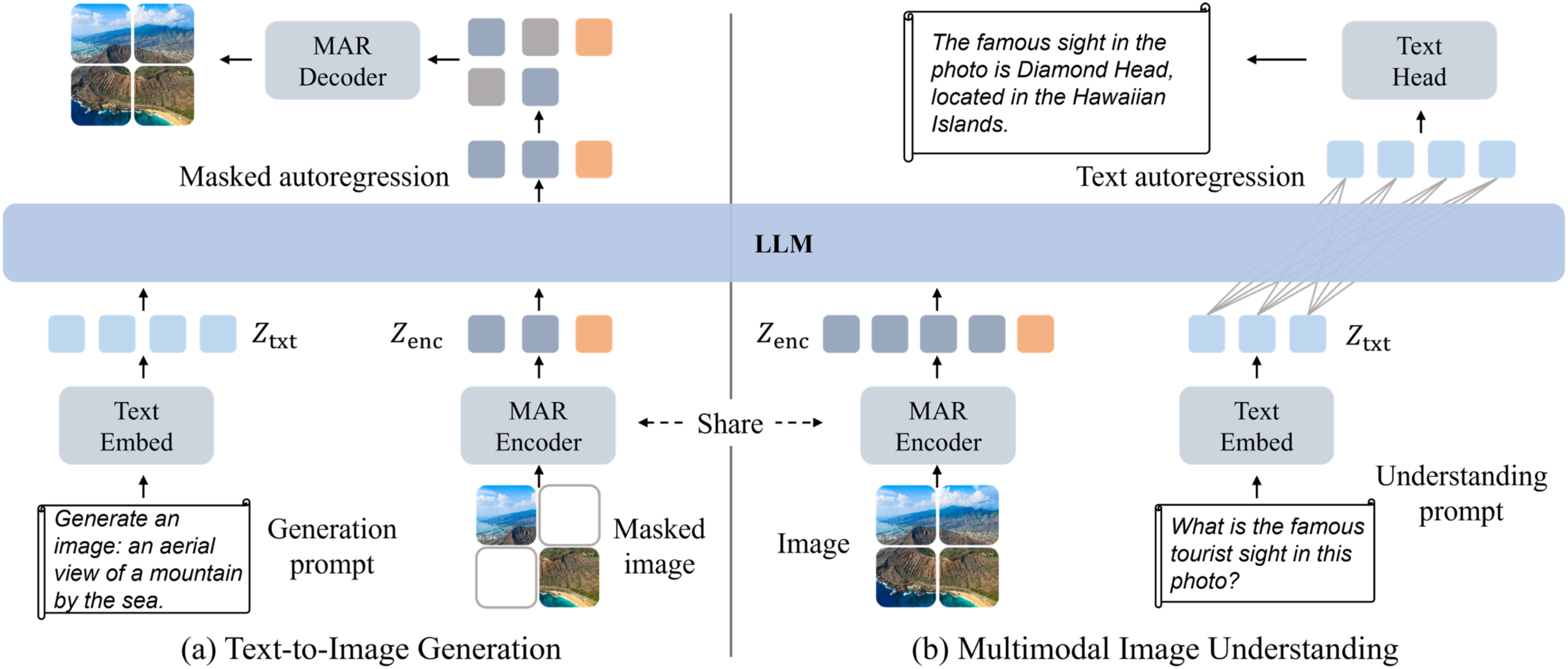
MAR representation is shared for image understanding and generation.

Outline

- Background
- Introduction
- **Method**
- Results

Method

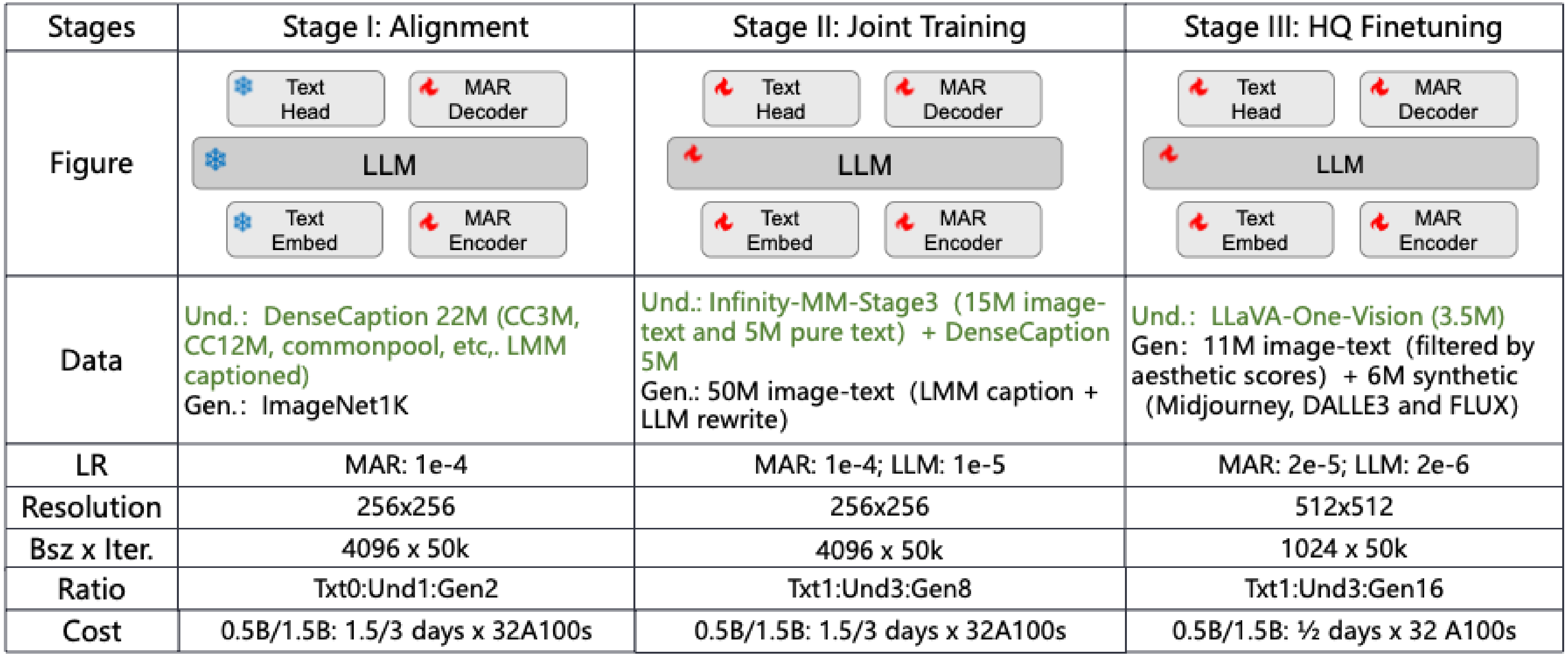
- **Harmon:** harmonizing visual representations for und. & gen.


MAR Decoder
Masked autoregression
LLM
Z_{txt}
Z_{enc}
Text Embed
MAR Encoder
Share
Generate an image: an aerial view of a mountain by the sea.
Generation prompt
Masked image
(a) Text-to-Image Generation
The famous sight in the photo is Diamond Head, located in the Hawaiian Islands.
Text Head
Text autoregression
Z_{enc}
Z_{txt}
MAR Encoder
Text Embed
Understanding prompt
Image
What is the famous tourist sight in this photo?
(b) Multimodal Image Understanding

The overview of our method: *Harmon.*

Method

- **Harmon:** harmonizing visual representations for und. & gen.



Stages	Stage I: Alignment	Stage II: Joint Training	Stage III: HQ Finetuning
Figure	❄ Text Head 🔥 MAR Decoder ❄ LLM ❄ Text Embed 🔥 MAR Encoder	🔥 Text Head 🔥 MAR Decoder 🔥 LLM 🔥 Text Embed 🔥 MAR Encoder	🔥 Text Head 🔥 MAR Decoder 🔥 LLM 🔥 Text Embed 🔥 MAR Encoder
Data	Und.: DenseCaption 22M (CC3M, CC12M, commonpool, etc,. LMM captioned) Gen.: ImageNet1K	Und.: Infinity-MM-Stage3 (15M image-text and 5M pure text) + DenseCaption 5M Gen.: 50M image-text (LMM caption + LLM rewrite)	Und.: LLaVA-One-Vision (3.5M) Gen: 11M image-text (filtered by aesthetic scores) + 6M synthetic (Midjourney, DALLE3 and FLUX)
LR	MAR: 1e-4	MAR: 1e-4; LLM: 1e-5	MAR: 2e-5; LLM: 2e-6
Resolution	256x256	256x256	512x512
Bsz x Iter.	4096 x 50k	4096 x 50k	1024 x 50k
Ratio	Txt0:Und1:Gen2	Txt1:Und3:Gen8	Txt1:Und3:Gen16
Cost	0.5B/1.5B: 1.5/3 days x 32A100s	0.5B/1.5B: 1.5/3 days x 32A100s	0.5B/1.5B: ½ days x 32 A100s

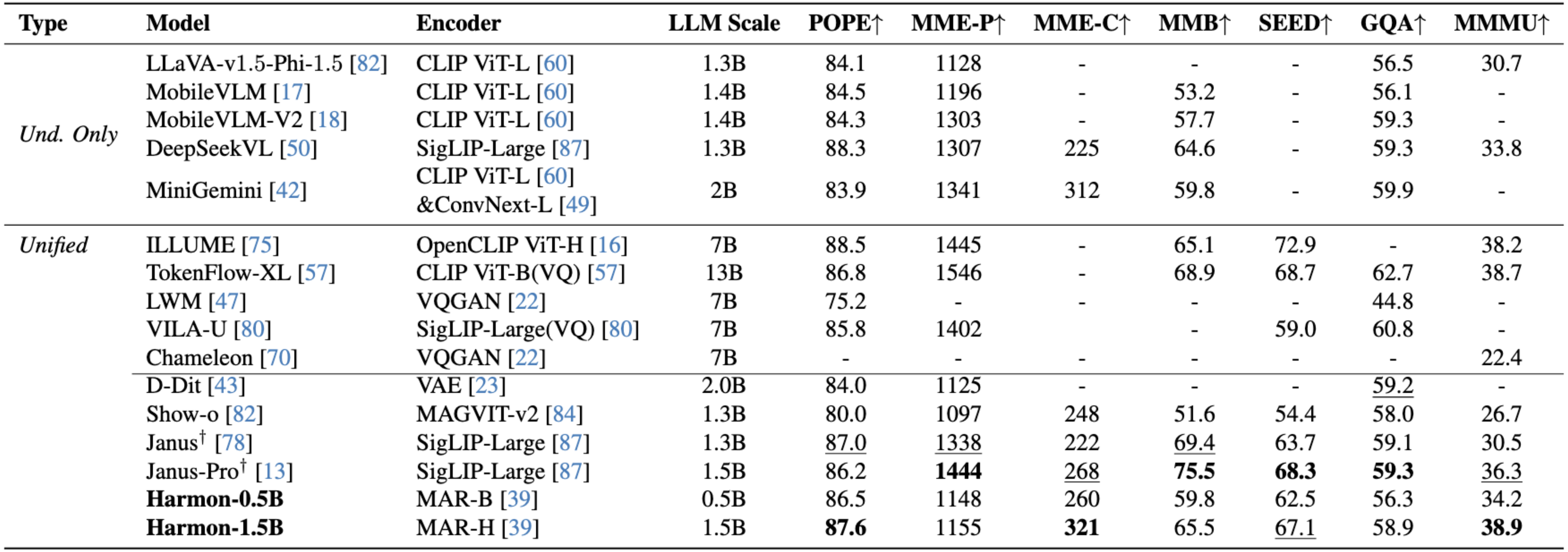
A three-stage training paradigm.

Outline

- Background
- Introduction
- Method
- Results

Results

- Image Understanding

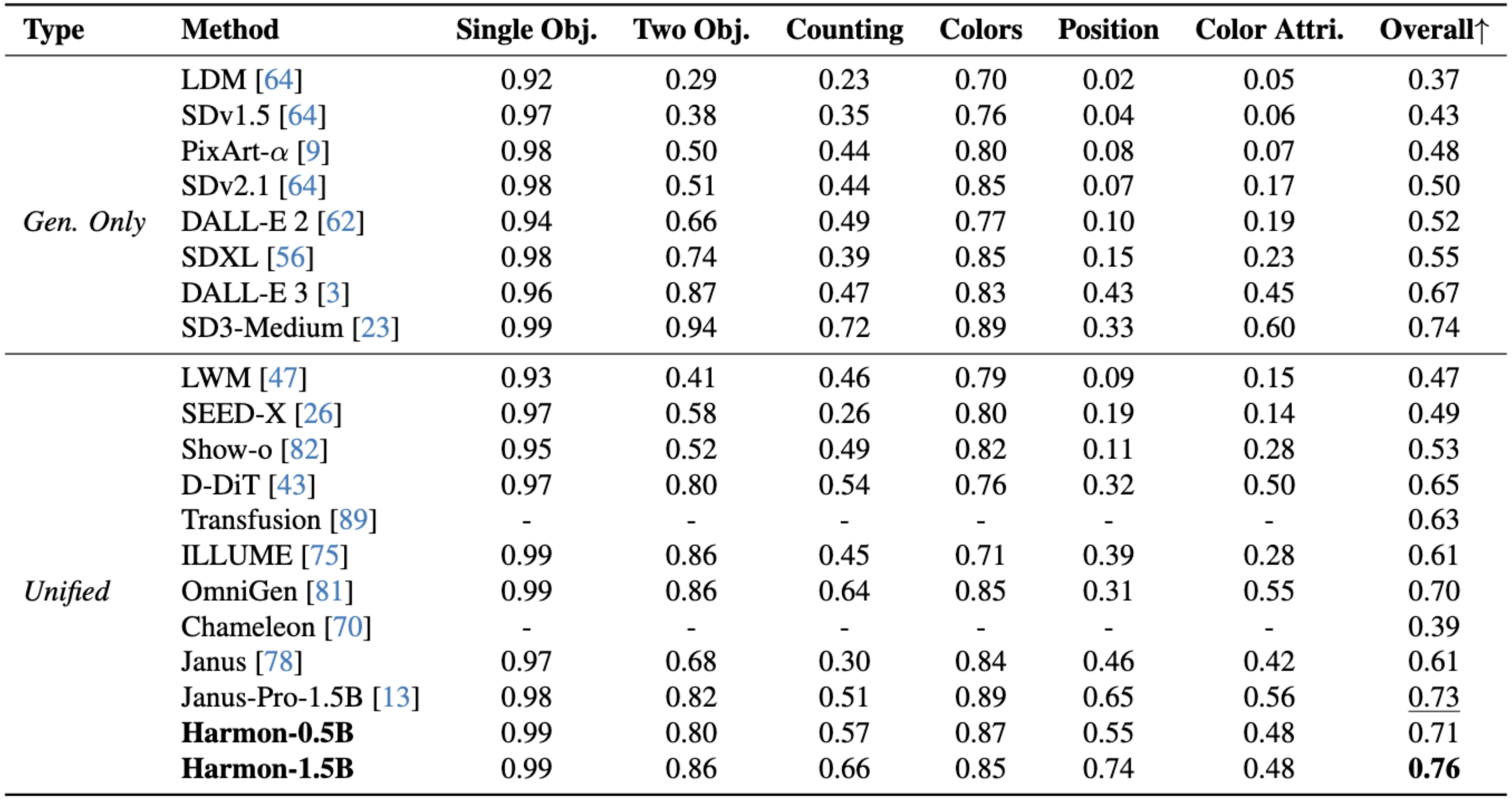


Type	Model	Encoder	LLM Scale	POPE↑	MME-P↑	MME-C↑	MMB↑	SEED↑	GQA↑	MMMU↑
Und. Only	LLaVA-v1.5-Phi-1.5 [82]	CLIP ViT-L [60]	1.3B	84.1	1128	-	-	-	56.5	30.7
	MobileVLM [17]	CLIP ViT-L [60]	1.4B	84.5	1196	-	53.2	-	56.1	-
	MobileVLM-V2 [18]	CLIP ViT-L [60]	1.4B	84.3	1303	-	57.7	-	59.3	-
	DeepSeekVL [50]	SigLIP-Large [87]	1.3B	88.3	1307	225	64.6	-	59.3	33.8
	MiniGemini [42]	CLIP ViT-L [60] &ConvNext-L [49]	2B	83.9	1341	312	59.8	-	59.9	-
Unified	ILLUME [75]	OpenCLIP ViT-H [16]	7B	88.5	1445	-	65.1	72.9	-	38.2
	TokenFlow-XL [57]	CLIP ViT-B(VQ) [57]	13B	86.8	1546	-	68.9	68.7	62.7	38.7
	LWM [47]	VQGAN [22]	7B	75.2	-	-	-	-	44.8	-
	VILA-U [80]	SigLIP-Large(VQ) [80]	7B	85.8	1402		-	59.0	60.8	-
	Chameleon [70]	VQGAN [22]	7B	-	-	-	-	-	-	22.4
	D-Dit [43]	VAE [23]	2.0B	84.0	1125	-	-	-	59.2	-
	Show-o [82]	MAGVIT-v2 [84]	1.3B	80.0	1097	248	51.6	54.4	58.0	26.7
	Janus† [78]	SigLIP-Large [87]	1.3B	87.0	1338	222	69.4	63.7	59.1	30.5
	Janus-Pro† [13]	SigLIP-Large [87]	1.5B	86.2	**1444**	268	**75.5**	**68.3**	**59.3**	36.3
	Harmon-0.5B	MAR-B [39]	0.5B	86.5	1148	260	59.8	62.5	56.3	34.2
	Harmon-1.5B	MAR-H [39]	1.5B	**87.6**	1155	**321**	65.5	67.1	58.9	**38.9**

Evaluation results on multimodal question-answering benchmarks.

Results

- Image Generation

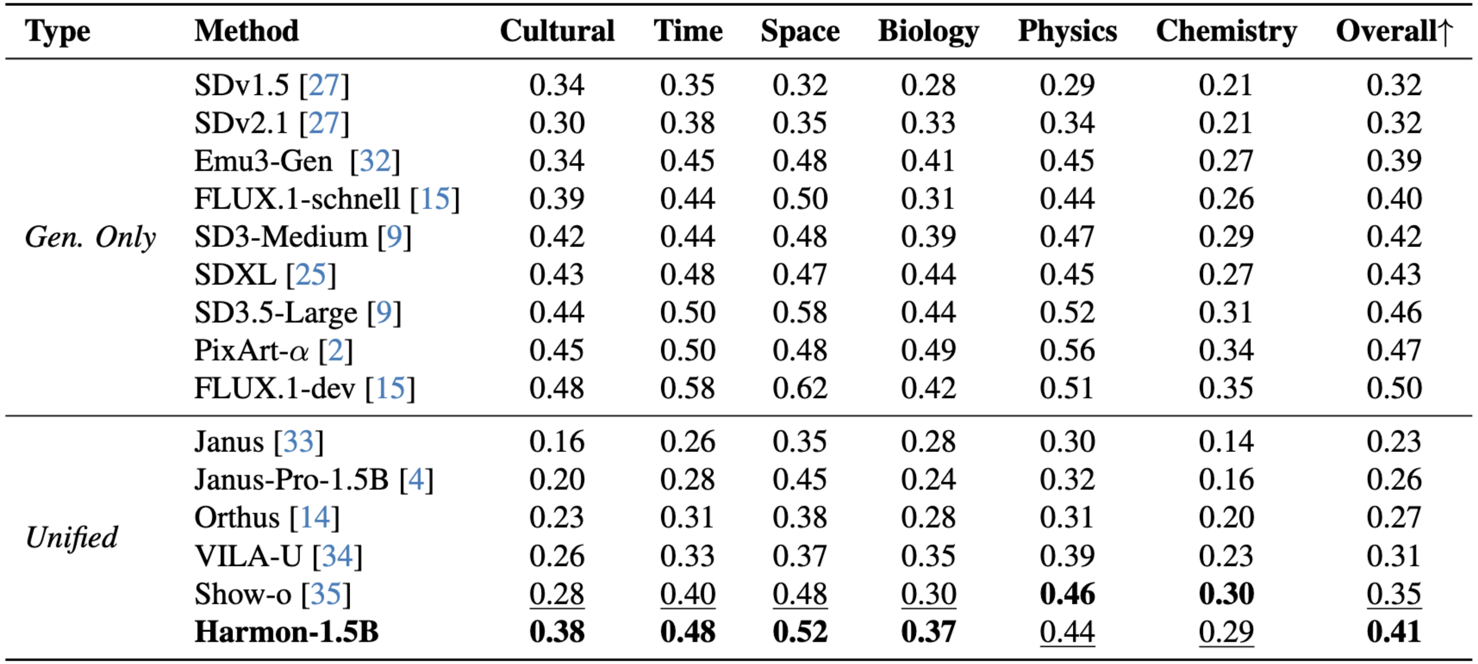


Type	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	LDM [64]	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [64]	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-α [9]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [64]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [62]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	SDXL [56]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [3]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [23]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Unified	LWM [47]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X [26]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o [82]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	D-DiT [43]	0.97	0.80	0.54	0.76	0.32	0.50	0.65
	Transfusion [89]	-	-	-	-	-	-	0.63
	ILLUME [75]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
	OmniGen [81]	0.99	0.86	0.64	0.85	0.31	0.55	0.70
	Chameleon [70]	-	-	-	-	-	-	0.39
	Janus [78]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Janus-Pro-1.5B [13]	0.98	0.82	0.51	0.89	0.65	0.56	0.73
	Harmon-0.5B	0.99	0.80	0.57	0.87	0.55	0.48	0.71
	Harmon-1.5B	0.99	0.86	0.66	0.85	0.74	0.48	**0.76**

Text-to-Image generation on GenEval benchmark.

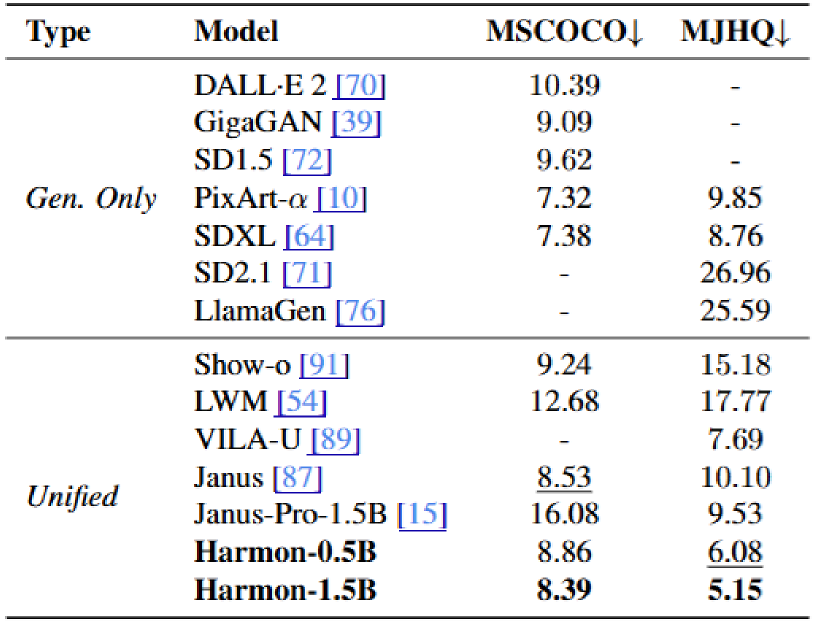
Results

- Image Generation



Type	Method	Cultural	Time	Space	Biology	Physics	Chemistry	Overall↑
Gen. Only	SDv1.5 [27]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
	SDv2.1 [27]	0.30	0.38	0.35	0.33	0.34	0.21	0.32
	Emu3-Gen [32]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
	FLUX.1-schnell [15]	0.39	0.44	0.50	0.31	0.44	0.26	0.40
	SD3-Medium [9]	0.42	0.44	0.48	0.39	0.47	0.29	0.42
	SDXL [25]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
	SD3.5-Large [9]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
	PixArt-α [2]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
	FLUX.1-dev [15]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified	Janus [33]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
	Janus-Pro-1.5B [4]	0.20	0.28	0.45	0.24	0.32	0.16	0.26
	Orthus [14]	0.23	0.31	0.38	0.28	0.31	0.20	0.27
	VILA-U [34]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
	Show-o [35]	0.28	0.40	0.48	0.30	**0.46**	**0.30**	0.35
	Harmon-1.5B	**0.38**	**0.48**	**0.52**	**0.37**	0.44	0.29	**0.41**

Text-to-Image generation on WISE benchmark.



Type	Model	MSCOCO↓	MJHQ↓
Gen. Only	DALL·E 2 [70]	10.39	-
	GigaGAN [39]	9.09	-
	SD1.5 [72]	9.62	-
	PixArt-α [10]	7.32	9.85
	SDXL [64]	7.38	8.76
	SD2.1 [71]	-	26.96
	LlamaGen [76]	-	25.59
Unified	Show-o [91]	9.24	15.18
	LWM [54]	12.68	17.77
	VILA-U [89]	-	7.69
	Janus [87]	8.53	10.10
	Janus-Pro-1.5B [15]	16.08	9.53
	Harmon-0.5B	8.86	6.08
	Harmon-1.5B	**8.39**	**5.15**

Text-to-Image generation on MSCOCO and MJHQ. The metric is FID.

Results

- Ablation study

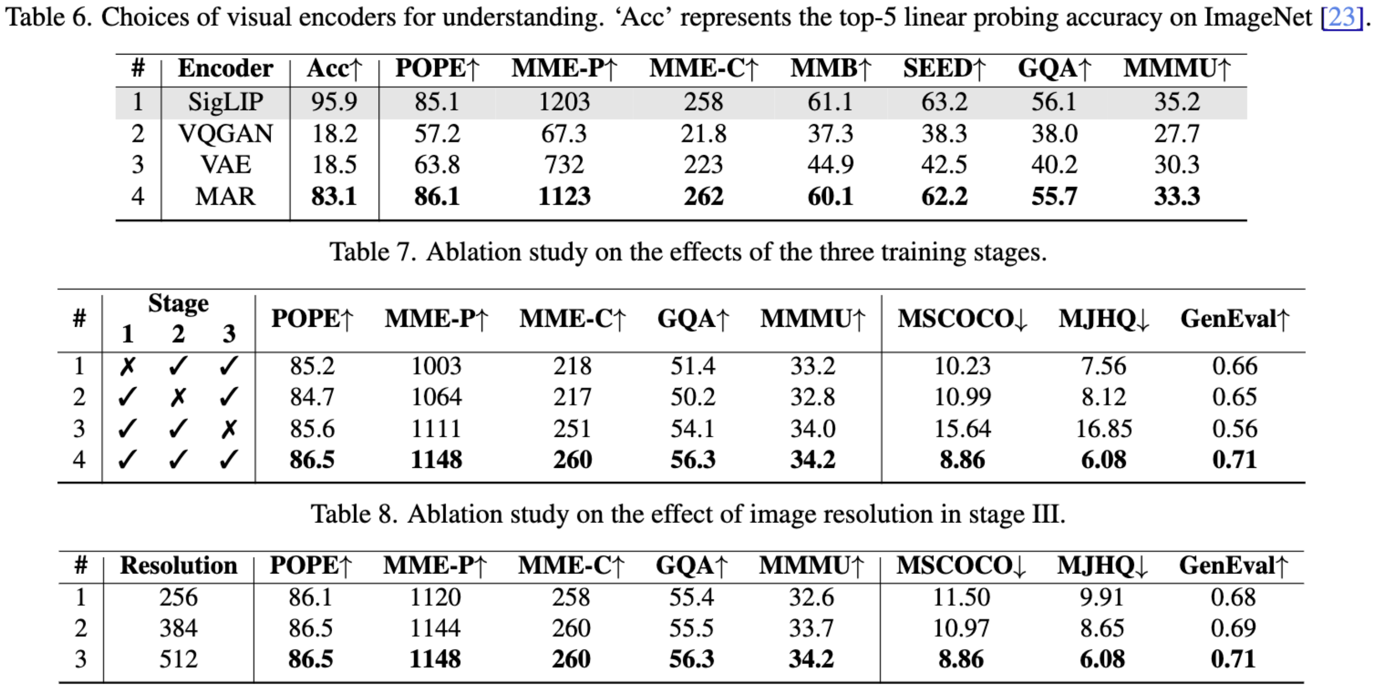


Table 6. Choices of visual encoders for understanding. 'Acc' represents the top-5 linear probing accuracy on ImageNet [23].

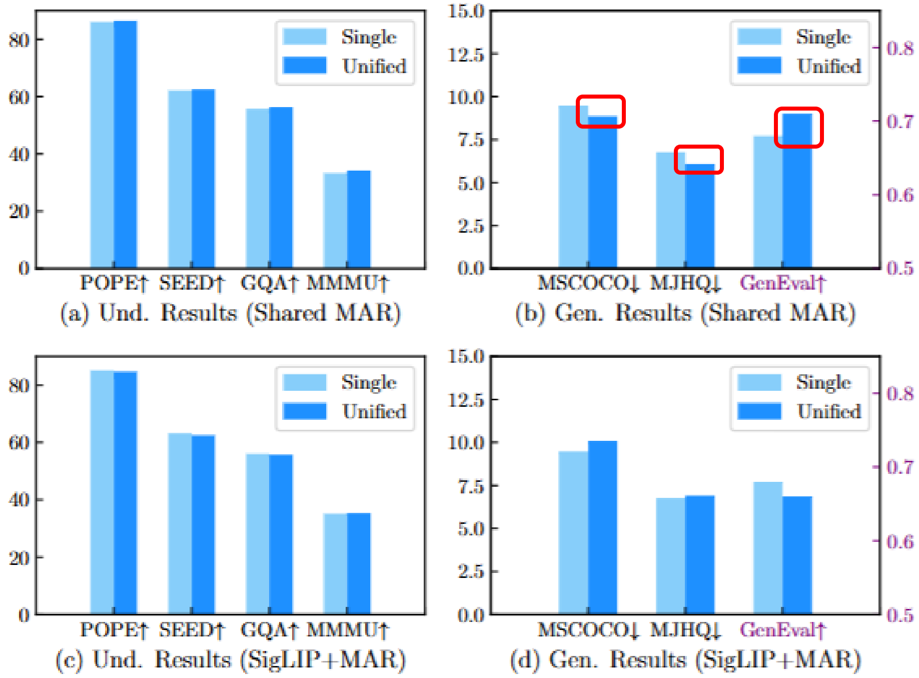
#	Encoder	Acc↑	POPE↑	MME-P↑	MME-C↑	MMB↑	SEED↑	GQA↑	MMMU↑
1	SigLIP	95.9	85.1	1203	258	61.1	63.2	56.1	35.2
2	VQGAN	18.2	57.2	67.3	21.8	37.3	38.3	38.0	27.7
3	VAE	18.5	63.8	732	223	44.9	42.5	40.2	30.3
4	MAR	**83.1**	**86.1**	**1123**	**262**	**60.1**	**62.2**	**55.7**	**33.3**

Table 7. Ablation study on the effects of the three training stages.

#	Stage 1	Stage 2	Stage 3	POPE↑	MME-P↑	MME-C↑	GQA↑	MMMU↑	MSCOCO↓	MJHQ↓	GenEval↑
1	✗	✓	✓	85.2	1003	218	51.4	33.2	10.23	7.56	0.66
2	✓	✗	✓	84.7	1064	217	50.2	32.8	10.99	8.12	0.65
3	✓	✓	✗	85.6	1111	251	54.1	34.0	15.64	16.85	0.56
4	✓	✓	✓	**86.5**	**1148**	**260**	**56.3**	**34.2**	**8.86**	**6.08**	**0.71**

Table 8. Ablation study on the effect of image resolution in stage III.

#	Resolution	POPE↑	MME-P↑	MME-C↑	GQA↑	MMMU↑	MSCOCO↓	MJHQ↓	GenEval↑
1	256	86.1	1120	258	55.4	32.6	11.50	9.91	0.68
2	384	86.5	1144	260	55.5	33.7	10.97	8.65	0.69
3	512	**86.5**	**1148**	**260**	**56.3**	**34.2**	**8.86**	**6.08**	**0.71**



(a) Und. Results (Shared MAR)

(b) Gen. Results (Shared MAR)

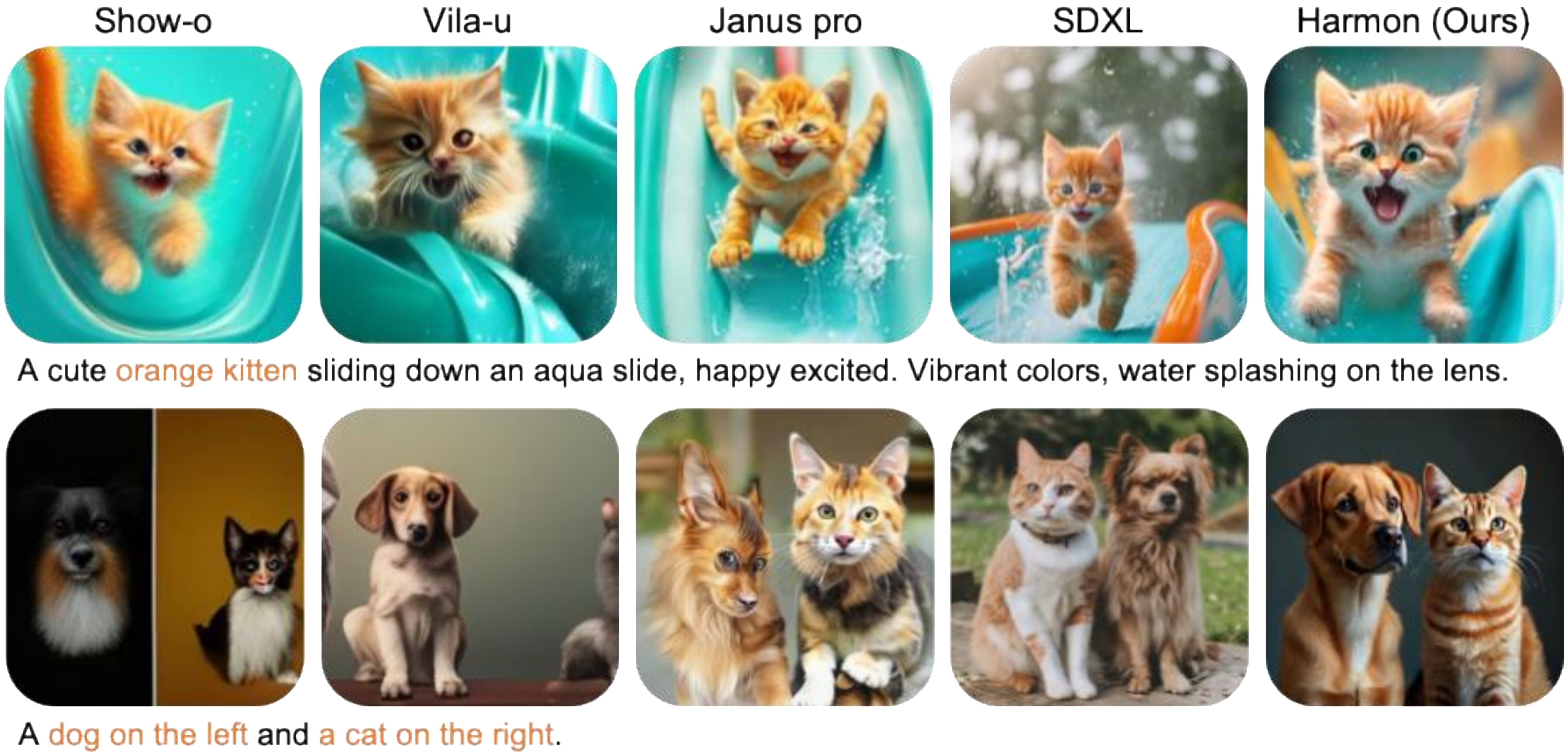
(c) Und. Results (SigLIP+MAR)

(d) Gen. Results (SigLIP+MAR)

Und. improves Gen.

Results

- Visualization: text-to-image generation

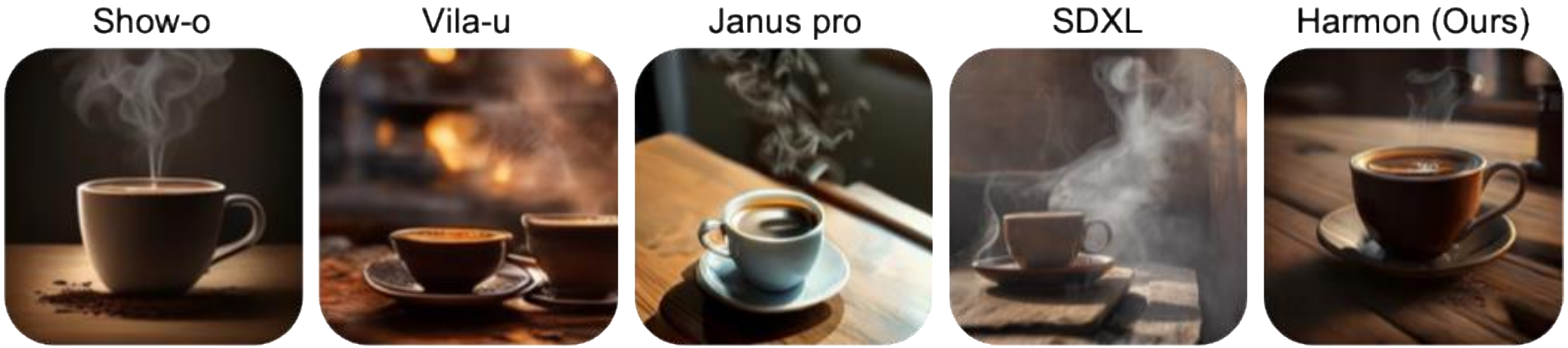


A cute orange kitten sliding down an aqua slide, happy excited. Vibrant colors, water splashing on the lens.

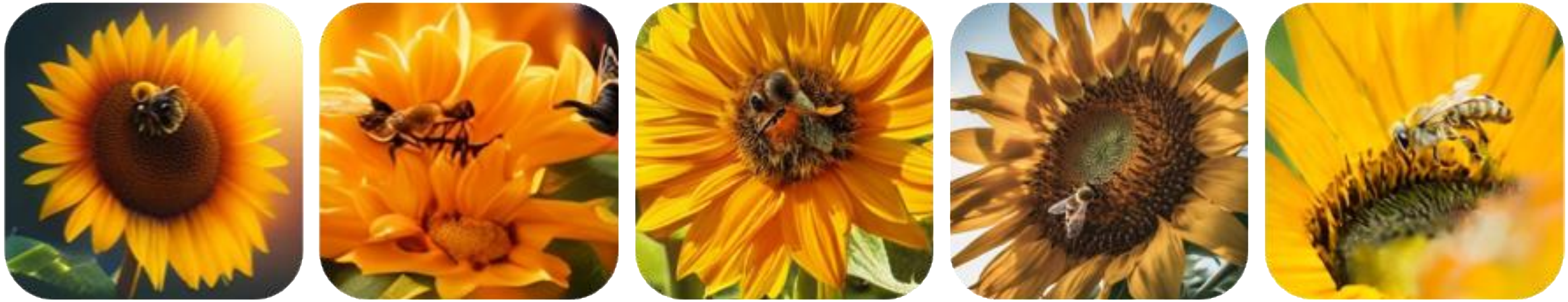
A dog on the left and a cat on the right.

Results

- Visualization: text-to-image generation



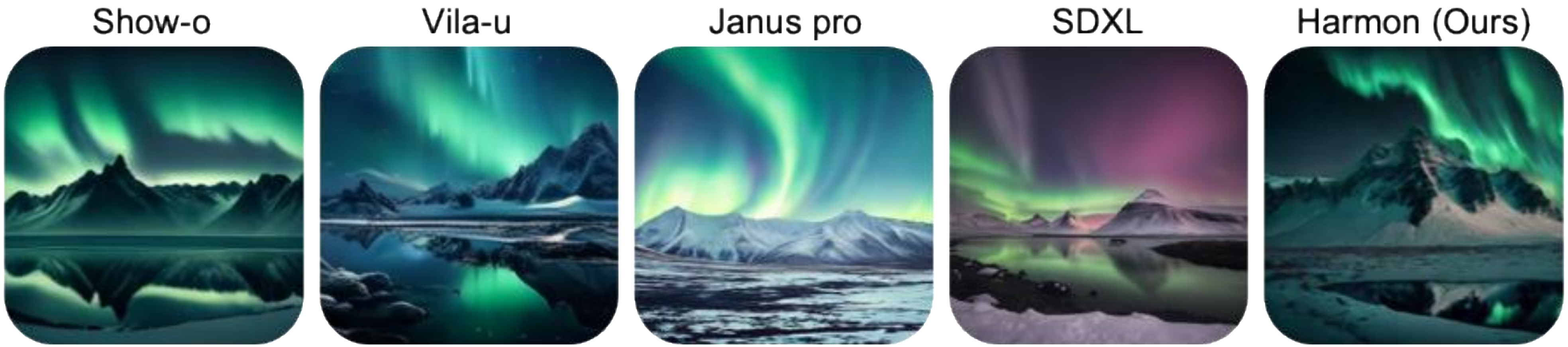
A steaming cup of coffee on a wooden table.



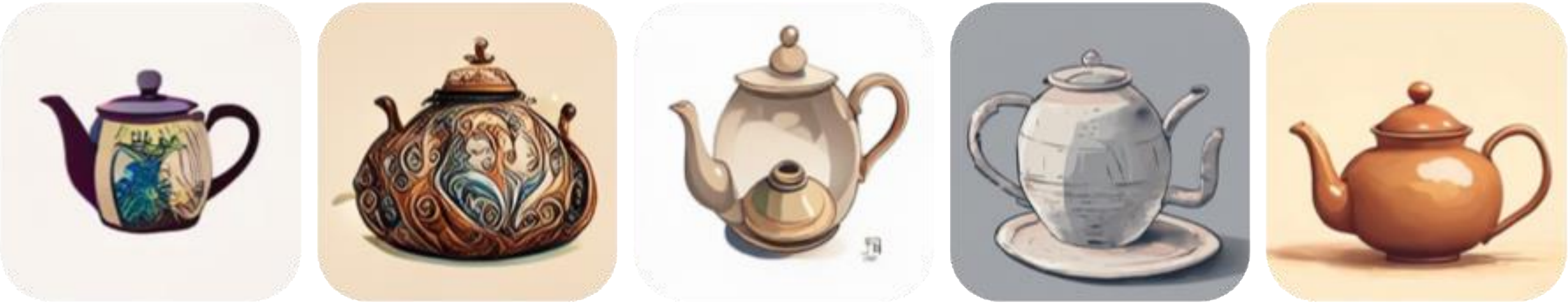
Capture a close-up shot of a vibrant sunflower in full bloom, with a honeybee perched on its petals, its delicate wings catching the sunlight.

Results

- Visualization: text-to-image generation



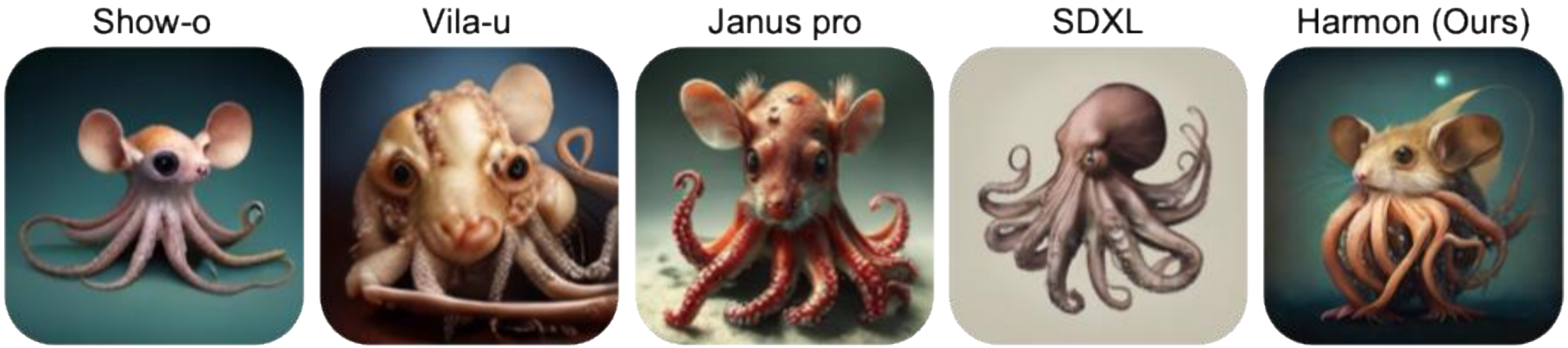
A realistic landscape shot of the Northern Lights dancing over a snowy mountain range in Iceland.



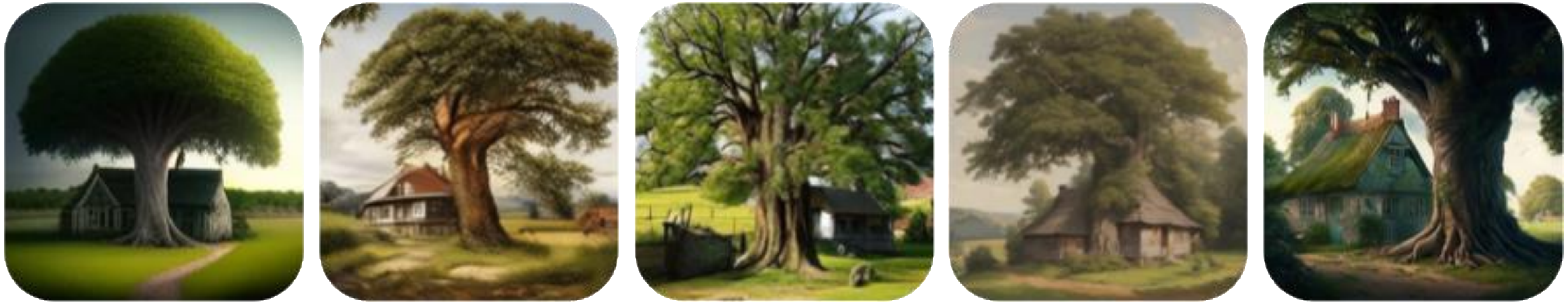
An illustration of a teapot.

Results

- Visualization: text-to-image generation



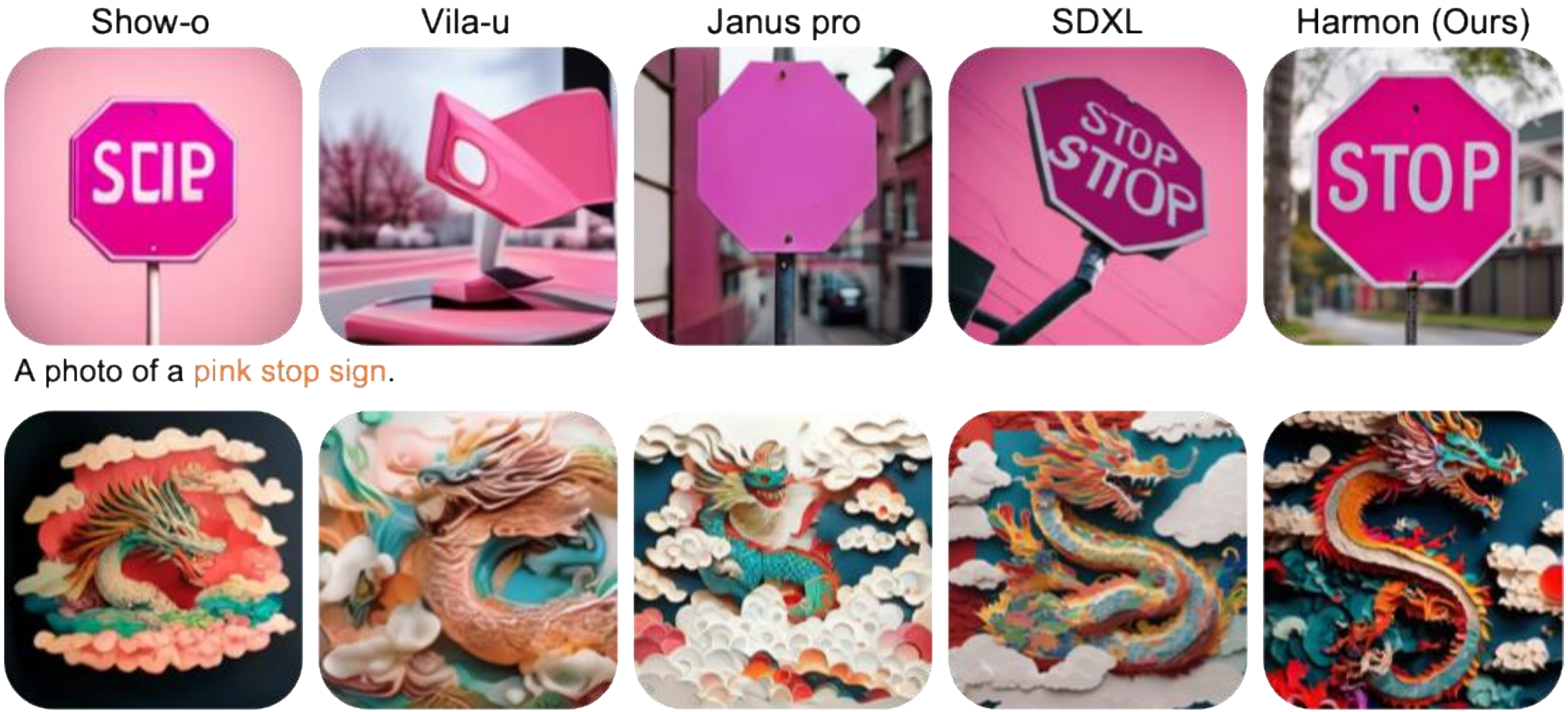
An image of an animal half mouse half octopus.



A big tree beside a rural cottage.

Results

- Visualization: text-to-image generation

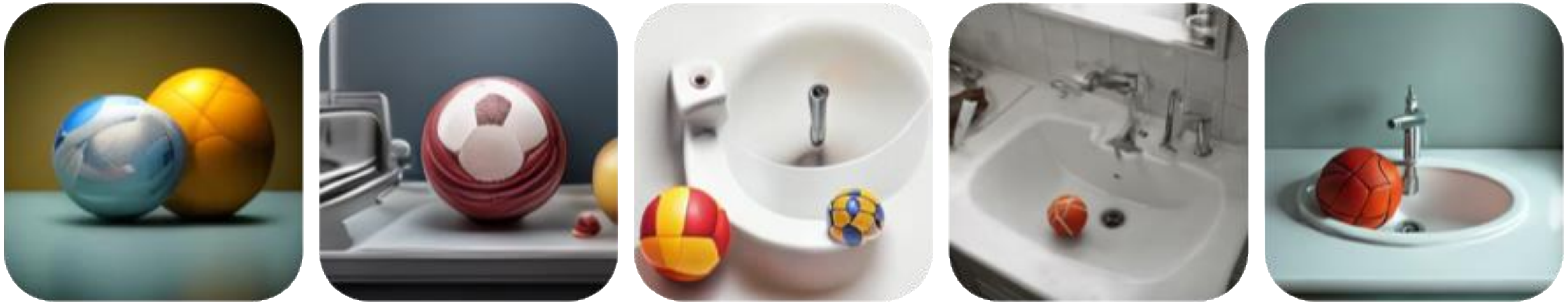


A photo of a pink stop sign.

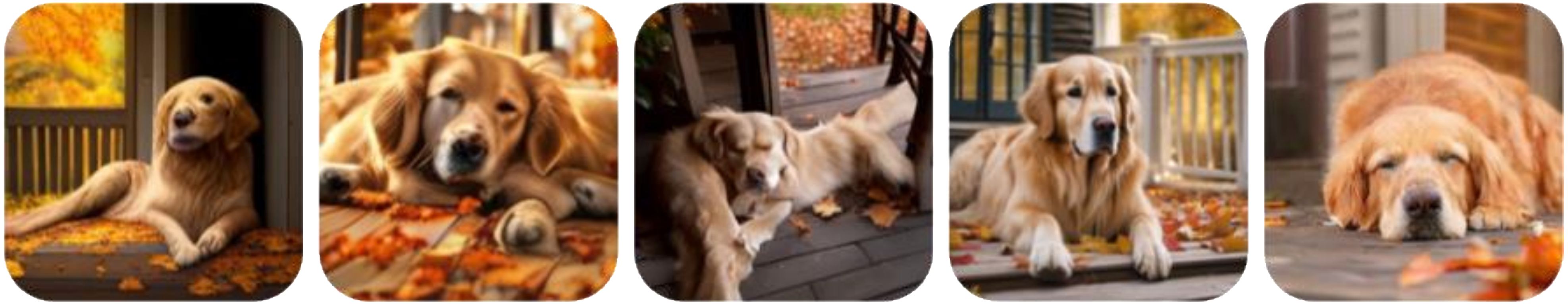
Paper artwork, layered paper, colorful Chinese dragon surrounded by clouds.

Results

- Visualization: text-to-image generation



A photo of a sink and a sports ball.



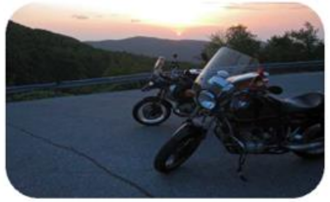
A golden retriever lying peacefully on a wooden porch, with autumn leaves scattered around.

Results

- Visualization: image understanding

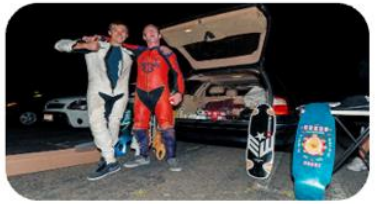
User: How many motorcycles are in the picture?
Model: There are two motorcycles in the picture.

User: How many people are on each motorcycle?
Model: There are no people on the motorcycles.



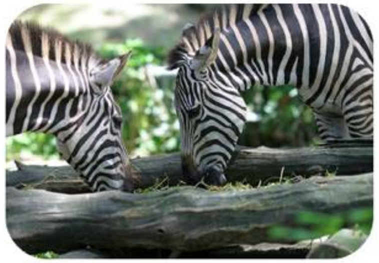
User: What is the color of the far right board?
Model: The color of the far right board is blue.

User: How many men are in the picture?
Model: There are two men in the picture.



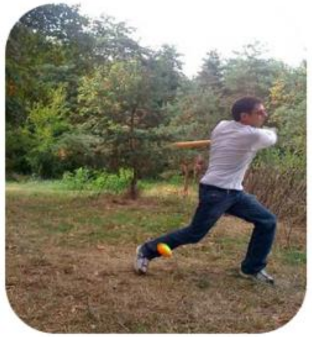
User: What animals are shown here?
Model: The animals shown in the image are zebras, specifically identified by their distinctive black and white striped coats.

User: What color are they?
Model: The zebras in the image are primarily black and white.



User: Is the man's legs together or apart?

Model: The man's legs are apart, with one leg extended forward and the other bent at the knee



Results

- Visualization: image understanding

User: Describe this scene in detail

Model: The image captures the breathtaking beauty of the Twelve Apostles, a famous landmark located along the coast of Victoria, Australia. These limestone stacks are formed from the remains of ancient forests that were subjected to powerful wave action over millions of years. The stacks stand majestically against the backdrop of the vast ocean, their unique shapes and sizes creating a striking contrast with the surrounding water. The colors in the image range from the deep blues of the ocean to the lighter hues of the sandy beach, adding depth and dimension to the scene. Above, the sky is filled with clouds, suggesting an overcast day but also enhancing the dramatic effect of the landscape below. This image encapsulates not only the physical attributes of this natural wonder but also the sense of awe it inspires in those who visit or view

User: Is she cooking bread?
Model: Yes, she is cooking bread. The image shows a person using an oven to bake several loaves of bread on a baking tray.

User: What room is this?
Model: The room in the image is a kitchen, which can be inferred from several clues. The presence of a stovetop with burners and control knobs indicates that this is an area where cooking takes place. Additionally, the tiled wall is commonly used for kitchens due to its durability and ease of cleaning. The utensils hanging on the wall further support this identification as they are typical kitchen items. Kitchens often have such appliances and fixtures designed for food preparation and cooking activities.

Thanks!