Supplementary Material of 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 Research
[3] The Chinese University of Hong Kong [4] SenseTime Research and Tetras.AI [5] SenseTime Research
size001@e.ntu.edu.sg {wei.l,ccloy}@ntu.edu.sg

S1. MAR

In this section, we provide more details of MAR [17].

Model Details. The MAR models in [17] follow the encoder-decoder architecture of MAE [12], and are trained on ImageNet1K [6] for image generation. Class embeddings are fed into the MAR encoder for class-conditional generation. An extra null embedding representing an empty class is also included for unconditional generation. In Harmon, we discard class embeddings and only use the null embeddings in MAR's forward pass (referred to as buffer embeddings in the main text). As a generation model, MAR follows the common practice [24–26] to compress images into VAE latents before feeding to MAR' encoder. For brevity, we omit the VAE part in our illustration of Harmon.

Potential for Understanding & Generation. We provide more visualization results in Figure S1 to unveil the MAR's potential for both visual understanding and generation. The feature activations in the second row of Figure S1 indicate that the MAR encoder has grasped essential visual concepts in its generative training. Then we map the encoder features back to image contents using the MAR decoder. It is noteworthy this operation is performed in a zero-shot manner as the MAR is trained for predicting unseen patches instead of pixel-level recovery. The results in the third row of Figure S1 suggest the MAR encoder's representation also contains intrinsic imagery features that are necessary for visual generation.

S2. Training Data

We provide details of out training data, including data sources and re-captioning processes.

S2.1. Image Understanding

Stage I. The 22M images with dense captions in stage I are sourced from LLaVA-ReCap-CC3M [16], PixelProse [29], DenseFusion [18] and the pre-training dataset of MiniGemini [19] and ShareGPT4V [3]. The dense captions in LLaVA-ReCap-CC3M are generated by LLaVA-NeXT-34B [16]. The PixelProse dataset comprises 16M images from CommonPool [10], CC12M [1] and RedCaps [7], which are re-captioned by Gemini-1.0-Pro-Vision [30]. DenseFusion labels 1M images from LAION [28] using a trained caption engine.

Stage II. The 20M comprehensive instruction-tuning data in stage II are from the Infinity-MM-Stage3 [11]. And extra 5M dense-captioned images are randomly sampled from the 22M images in our stage I.

Stage III. In the high-quality fine-tuning stage, we directly use instruction-tuning data from LLaVA-One-Vision [16] for image understanding.

S2.2. Image Generation

Stage I. For class-conditional image generation in stage I, we use ImageNet1K [6] with 1.2M data samples, treating class names as image captions.

Stage II. For text-to-image generation, we first rewrite the 22M dense captions in stage I into shorter descriptions with Qwen2.5-7B-Instruct [36], using the following prompt:
"`Here is a detailed image description: <caption>. Rewrite it into a much shorter, vivid, and visually rich sentence (one or two sentences) that captures only the most essential elements and atmosphere of the scene. Ensure the description is concise, clear, and optimized for use with a text-to-image generation model.`"
Here, `<caption>` stands for the dense caption.

In addition, we import datasets specially collected for image generation, including PD12M [21], Megalith10M [20] and LAION-Aesthetics [5]. Like the prior

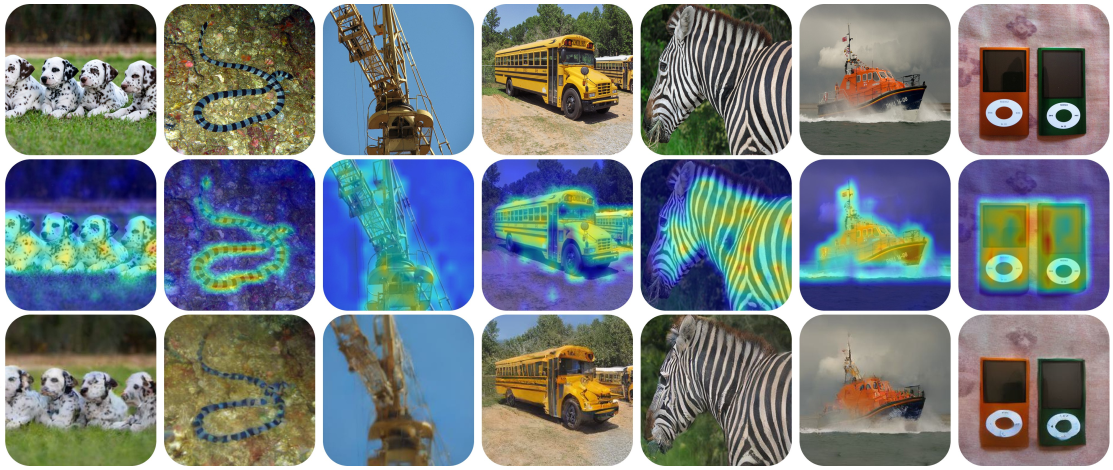


Figure S1. We visualize activations on MAR's feature maps in the second row, which reveal precise responses to visual concepts. In the third row, we observe that the features can also be mapped back to image pixels, indicating that the MAR features also comprise low-level image intrinsics.

22M dense caption data, the PD12M dataset is originally labelled with detailed image descriptions. Therefore, we also use Qwen2.5-7B-Instruct to re-write all the image descriptions with the prompt defined above. For Megalith10M, we directly use the short captions provided by [8]. For LAION-Aesthetics, we crawled 6M images using their urls and labelled them with precise generation prompts by Qwen2-VL-72B [31].

In total, we collect 50M data samples for training text-to-image generation in stage II.

Stage III. For high-quality text-to-generation, we apply an aesthetic prediction model [5] to score the 50M images in stage II. Only images with aesthetic scores beyond 6.5 are preserved. Further, we discard images with extreme height-width ratios. Finally, 10M images are selected for stage II. Additionally, we obtain 6M synthetic images from JourneyDB [23] and Text-to-Image-2M [13] to further enhance visual quality.

S3. More results

S3.1. Benchmark Results

We assess Harmon's ability to understand complex semantic and world knowledge using the WISE benchmark [22], where implicit prompts like "`Einstein's favorite musical instrument`" are provided. As shown in Table S3, Harmon archives the best performance among all compared unified models.

S3.2. Inference Speed

By default, we adopt 64 forward passes for generation, costing 10s/30s for Harmon-0.5B/1.5B on an A100 GPU. To speed up inference, we can reduce the forward steps to 16 (3s/8s) without an obvious performance drop on GenEval as shown in Table S2.

S3.3. Visualization

Qualitative Comparison. We provide qualitative comparison on text-to-image generation in Figure S2 and Figure S3. Here, we compare Harmon-1.5B with unified models including VILA-U [34], Show-o [35] and Janus-Pro [4](1.5B). Besides, we also include SDXL [25], an advanced expert model for visual generation. Harmon produces results comparable to SDXL in terms of visual quality, and exhibits better prompt-mage consistency. For example, SDXL fails to follow the position relations defined by 'A dog on the left and a cat on the right' in Figure S2.

More Gen. & Und. Results. We show more examples of Harmon-1.5B performing text-to-image generation in Figure S4 and multimodal understanding in Figure S5.

S4. Limitations

Despite promising results on both visual understanding and generation tasks, the current version of Harmon has the following limitations.

Model Scale. Our model scale is limited to 1.5B and we will further scale up the model size in the future.

Pre-training of MAR. The MAR models are originally pre-trained on the 1.2M data samples of ImageNet1K, which is orders of magnitude fewer than the billion-scale training of semantic encoders like CLIP and SigLIP. This gap in data scale hinders further improvement of Harmon in understanding tasks.

Table S1. Evaluation of text-to-image generation on WISE benchmark. *Gen. Only* stands for models trained for image generation only.

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]	<u>0.28</u>	<u>0.40</u>	<u>0.48</u>	<u>0.30</u>	**0.46**	**0.30**	<u>0.35</u>
	Harmon-1.5B	**0.38**	**0.48**	**0.52**	**0.37**	<u>0.44</u>	<u>0.29</u>	**0.41**

Table S2. Performance on GenEval for different inference steps.

#Steps	Harmon-1.5B	Harmon-0.5B
64	0.76	0.71
32	0.76	0.71
16	0.74	0.69
8	0.66	0.60
4	0.47	0.44

References

[1] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 1

[2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3, 4

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 1

[4] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2, 3, 4

[5] dclure. Laion-aesthetics-umap. https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap, 2022. 1, 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[7] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 1

[8] Caption Emporium. flickr-megalith-10m-internvl2-multi-caption. https://huggingface.co/datasets/CaptionEmporium/flickr-megalith-10m-internvl2-multi-caption, 2024. 2

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 3, 4

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 1

[11] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024. 1

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[13] jackyhate. text-to-image-2m. https://huggingface.co/datasets/jackyhate/text-to-image-2M, 2024. 2

[14] Siqi Kou, Jiachun Jin, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autore-

Table S3. Evaluation of text-to-image generation on WISE benchmark. *Gen. Only* stands for models trained for image generation only.

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]	<u>0.28</u>	<u>0.40</u>	<u>0.48</u>	<u>0.30</u>	**0.46**	**0.30**	<u>0.35</u>
	Harmon-1.5B	**0.38**	**0.48**	**0.52**	**0.37**	<u>0.44</u>	<u>0.29</u>	**0.41**

gressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*, 2024. 3, 4

[15] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3, 4

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[17] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024. 1

[18] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024. 1

[19] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 1

[20] madebyollin. Megalith-huggingface. https://huggingface.co/datasets/madebyollin/megalith-10m, 2024. 1

[21] Jordan Meyer, Nick Padgett, Cullen Miller, and Laura Exline. Public domain 12m: A highly aesthetic image-text dataset with novel governance mechanisms. *arXiv preprint arXiv:2410.23144*, 2024. 1

[22] 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. 2

[23] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023. 2

[24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[25] 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. 2, 3, 4

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

[29] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024. 1

[30] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[31] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

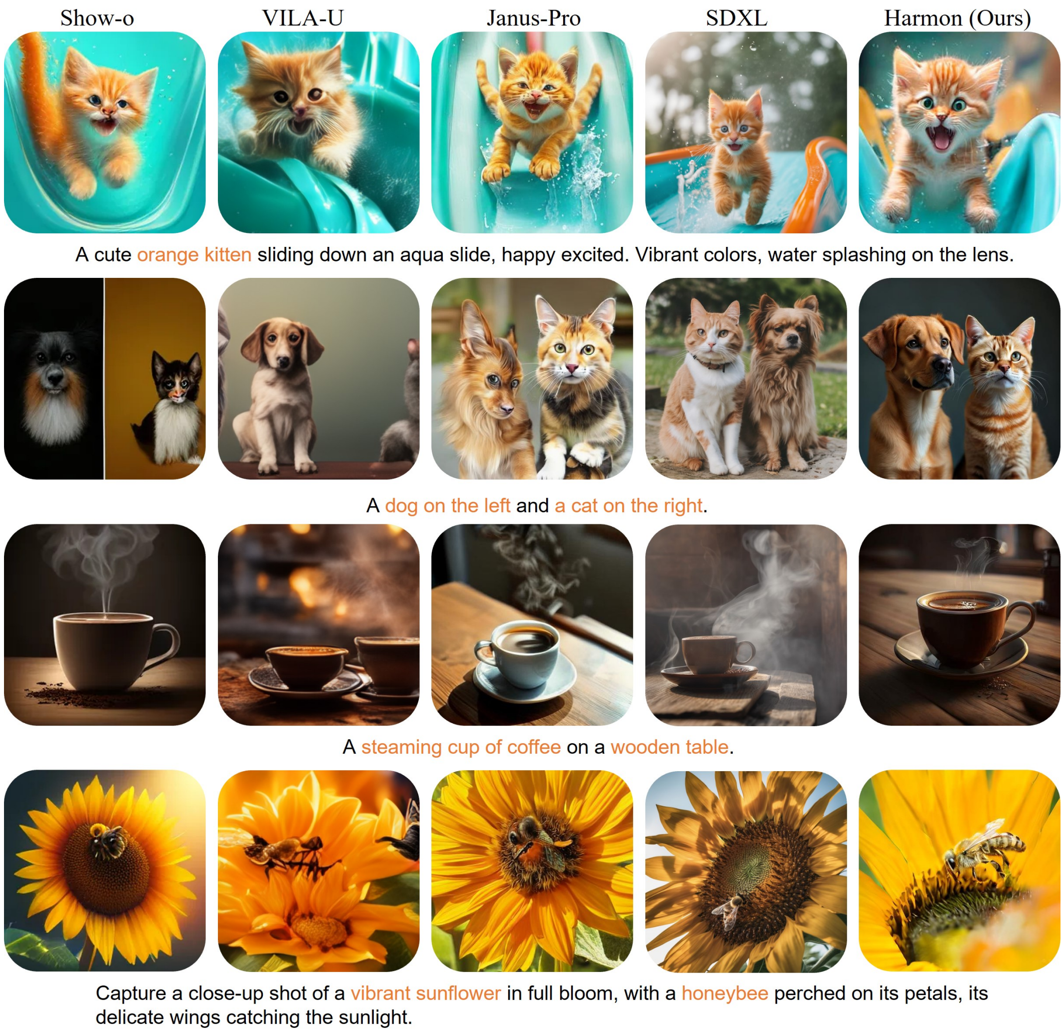


Figure S2. Qualitative comparison between Show-o-1.3B-512, VILA-U, Janus-Pro-1.5B and our Harmon-1.5B on text-to-image generation. The text below each image represents the generation prompt, with key terms guiding the generation highlighted in orange. Best viewed on screen.

[32] 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. 3, 4

[33] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 3, 4

[34] 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. 2, 3, 4

[35] 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. 2, 3, 4

[36] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1

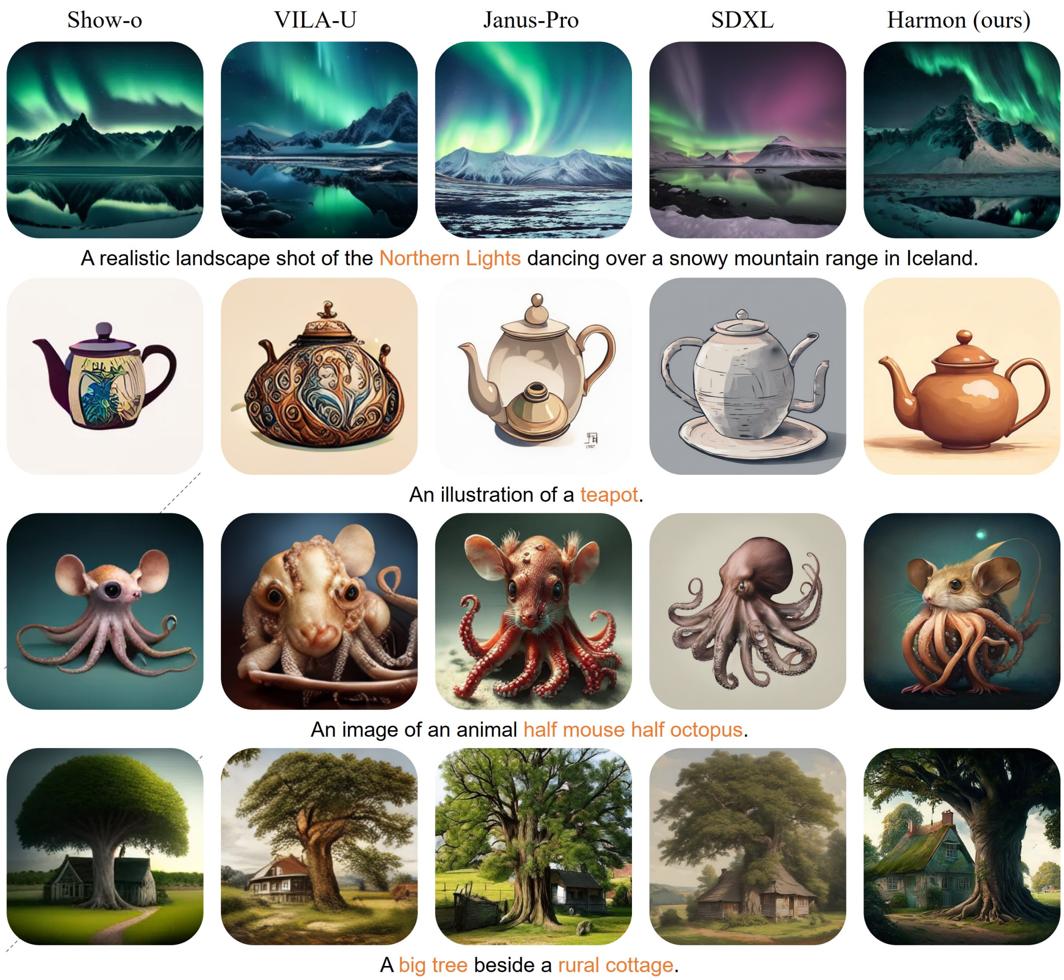


Figure S3. Qualitative comparison between Show-o, VILA-U, Janus-Pro (1.5B) and our Harmon (1.5B) on text-to-image generation. The text below each image represents the generation prompt, with key terms guiding the generation highlighted in orange. Best viewed on screen.

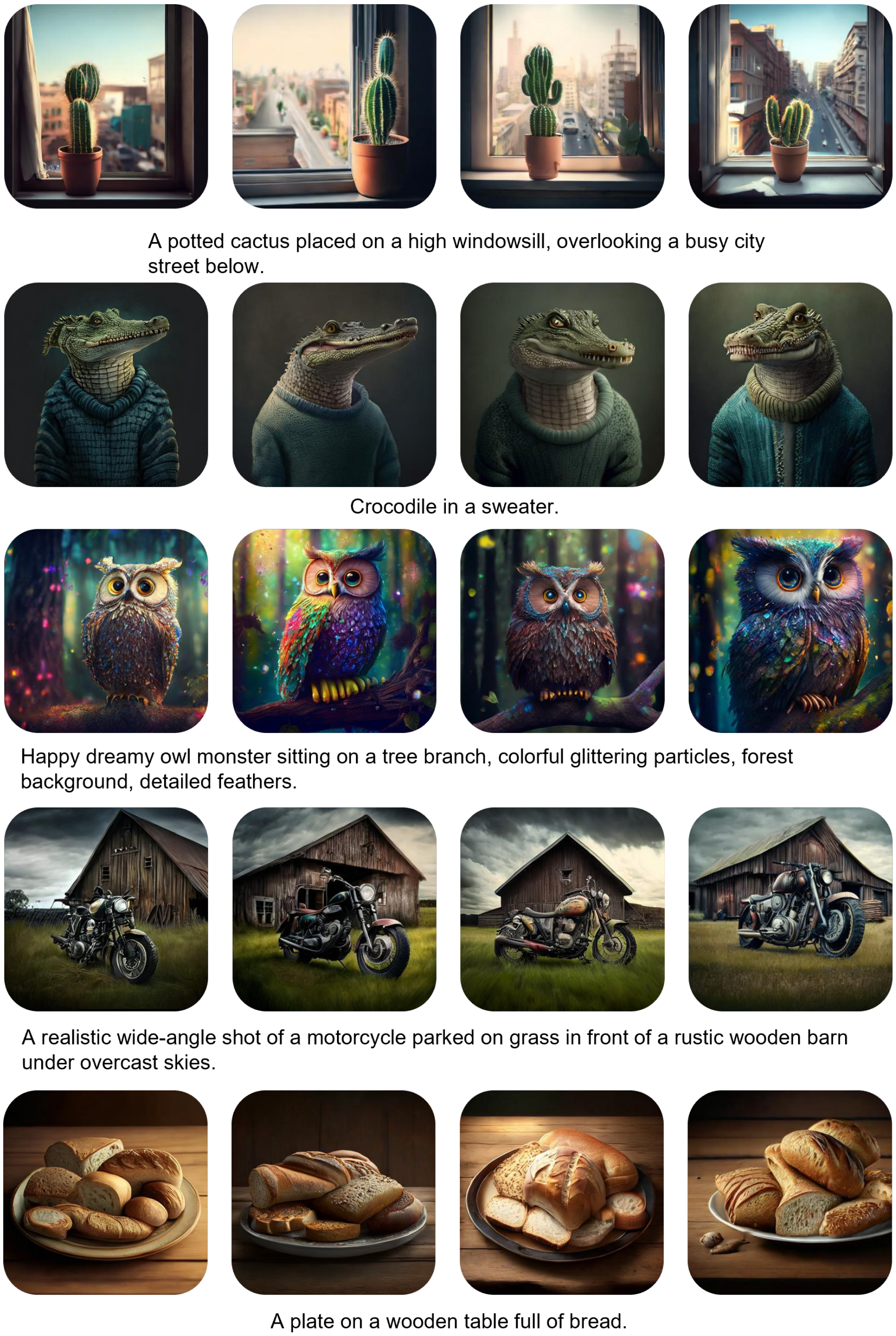


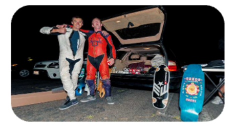
Figure S4. Text-to-image generation results by Harmon-1.5B. Our model is able to generate precise and diverse images based on text prompts.

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

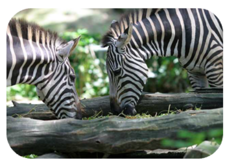
User: How many people are on each motorcycle?
Model: There are no people on the motorcycles.

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

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.

Figure S5. Examples of multimodal image understanding in visual question-answering format. The results are obtained by Harmon-1.5B.