MUSE-VL: Modeling Unified VLM through Semantic Discrete Encoding

Supplementary Material

Appendix

A. Additional Results

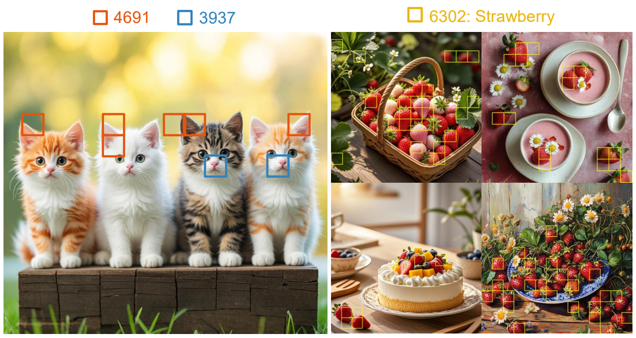

Figure 6. Visualization of semantic discrete codes. Rectangular boxes of the same color indicate that the corresponding semantic ID of these patches is the same. It can be observed that semantic ID can represent a semantic concept.

Visualization of Semantic Code Figure 6 shows the visualization of semantic encoding. We convert the image into discrete codes using the proposed SDE tokenizer, group the patches of the image according to their codes, and mark them with rectangular boxes. The left image indicates that the two IDs represent the cat's ears and the area near its nose, respectively. The right image visualizes the code that represents strawberries. The illustration demonstrates that the discrete codes extracted by the SDE tokenizer contain high-level semantic information, thus significantly enhancing the understanding capability (as shown in Table 6).

Image Reconstruction Figure 7 shows the comparison of the image reconstruction results with other semantic tokenizers, where the first column is the original image. We observe that methods like SEED [3] and LaViT [6] can only retain basic semantic information, but show significant differences in color, number of objects, and background compared to the original image. Emu2 [14] failed to accurately restore some details (the rectangular box in the figure). The proposed tokenizer explicitly integrates high-level semantic information and low-level information during the discretization process, so the reconstructed results perform better in preserving both the major objects and the details.

Visualization of VQA Figure 8 illustrates MUSE-VL's ability to tackle visual question-answering tasks. The model receives an image as its initial input, after which the user poses questions regarding the image. The results show that the Chameleon [15] and SEEDLLaMA [3] models make obvious errors in animal recognition and spatial localization. Additionally, Chameleon describes objects that were not present in the image, indicating hallucination issues. Compared with them, the results show the proposed model can accurately answer questions based on image information, demonstrating that the model has effective spatial localization and instruction-following capabilities.

Figure 7. Comparison of image reconstruction results with different methods [3, 6, 14]. The original image is in the first column, and SDE is the proposed tokenizer.

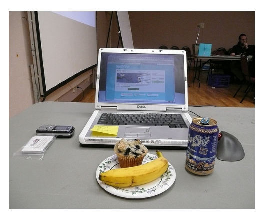

Figure 8. Comparison of results on the Visual Question Answering (VQA) task. The model is required to answer the user's questions based on the input image. The inaccurate parts of the response are highlighted in red.

Type	Method	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.
Gen. Only	DALL-E 2 [11]	0.52	0.94	0.66	0.49	0.77	0.1	0.19
	SDv1.5 [12]	0.43	0.97	0.38	0.35	0.76	0.04	0.06
	SDv2.1 [12]	0.50	0.98	0.51	0.44	0.85	0.07	0.17
	SDXL [10]	0.55	0.98	0.74	0.39	0.85	0.15	0.23
	PixArt-alpha [2]	0.48	0.98	0.5	0.44	0.8	0.08	0.07
	DALL-E 3 [1]	0.67 †	0.96	0.87	0.47	0.83	0.43	0.45
	LlamaGen [13]	0.32	0.71	0.34	0.21	0.58	0.07	0.04
Und. and Gen.	Chameleon [15]	0.39	-	-	-	-	-	-
	LWM [9]	0.47	0.93	0.41	0.46	0.79	0.09	0.15
	SEED-X [4]	0.49	0.97	0.58	0.26	0.8	0.19	0.14
	Show-o [20]	0.53	0.95	0.52	0.49	0.82	0.11	0.28
	Ours (7B)	0.53	0.99	0.65	0.44	0.73	0.18	0.17
	Ours (7B)	0.57 †	0.98	0.64	0.52	0.72	0.25	0.31

Table 8. Evaluation of text-to-image generation on the GenEval [5]. † result is with rewriting.

Table 9. Evaluation on TextVQA benchmark.

Method	Resolution	TextVQA
LLAVA 1.5 [8]	336	58.2
Janus [18]	384	50.7
VILA-U [19]	256	48.3
VILA-U [19]	384	60.8
EMU3 [17]	1024	64.7
SynerGen-VL [7]	Dynamic	67.5
MUSE-VL	256	52.8
MUSE-VL	384	61.3

Benchmarks of high-resolution benchmarks Table 9 presents the results of the commonly used TextVQA benchmark. This benchmark is highly relevant to OCR tasks and therefore requires high-resolution image understanding capabilities. It is worth noting that our current model does not yet include the training process of high-resolution images. We plan to support high-resolution input in future work.

B. Visual Generation Results

Table 8 shows the quantitative results of the text-to-image in GenEval [5] benchmark and compares them with other state-of-the-art generation models. We followed DALL-E 3 [1] to rewrite the prompts, making them more aligned with the dense captions in the training data. The results show that our model exhibits better performance than other unified models such as Chameleon [15] and SEED-X [4]. And it achieves performance close to the diffusion models. This indicates that our model has a strong image-text alignment capability.

C. Limitation and Future Work

Due to the limitations in the scale of training data and the resolution of generated images, our model has not surpassed the SOTA diffusion models in visual generation. In the future, we plan to further enhance the generation quality by expanding the scale of the training dataset for visual generation and using a more powerful image encoder [16]. Furthermore, exploring the native integration of AR and Diffusion to further enhance the quality of image generation and instruction following is both challenging and promising.

In this work, extensive experiments and evaluations have been conducted on multimodal understanding and text-to-image tasks, demonstrating that our model can effectively unify the modeling of textual and visual data. Moreover, the architecture of our model supports arbitrary sequences of images and text. The next step is to further expand the capabilities of MUSE-VL by incorporating interleaved image-text data and image-editing data during training.

References

[1] James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. Improving image generation with better captions, 2023. 2

[2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 2

[3] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 1

[4] 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. 2

[5] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: an object-focused framework for evaluating text-to-image alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[6] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Yadong Mu, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In *International Conference on Learning Representations*, 2024. 1

[7] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29767–29779, 2025. 2

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[9] 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. 2

[10] 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 *The Twelfth International Conference on Learning Representations*, 2024. 2

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2

[13] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 1

[15] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 2

[16] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 2

[17] 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. 2

[18] 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. 2

[19] 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

[20] 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