EFFICIENT MULTIMODAL GENERATION VIA REDUNDANCY-AWARE MIXTURE-OF-EXPERTS

Raman Dutt[♠], **Harleen Hanspal**[◇], **Petru-Daniel Tudosiu**[♣], **Alexander Black**[†], **Yongxin Yang**[♦], **Steven McDonagh**[♠], **Sarah Parisot**[‡]

[♠] The University of Edinburgh
[◇] Imperial College, London
[♣] Leonardo.AI
[†] University of Surrey
[♦] Queen Mary University of London
[‡]Microsoft Research, Cambridge

raman.dutt@ed.ac.uk, h.hanspal21@imperial.ac.uk, daniel.tudosiu@leonardo.ai
alexander@black.com, yongxin.yang@qmul.ac.uk, s.mcdonagh@ed.ac.uk
sarahparisot@microsoft.com

ABSTRACT

Multimodal foundation models are increasingly explored under diverse generation paradigms beyond classic next-token prediction. In this work, we study how autoregressive multimodal generation can be efficiently extended by exploiting latent capacity already present in models in the form of redundant parameters. We address the problem of augmenting pre-trained text-only LLMs with multimodal generative capabilities under two constraints: **(C1)** preserving original language generation performance, and **(C2)** maintaining a small parameter and data budget. Rather than introducing modality-specific modules, we leverage expert redundancy in Mixture-of-Experts (MoE) architectures as a source of latent capacity for learning a new modality. To prevent catastrophic forgetting, we apply Partial Low-Rank Adaptation (PLoRA) exclusively to tokens of the new modality, leaving text pathways unchanged. Through continual multimodal fine-tuning, our approach enables high-fidelity text-to-image generation while preserving original language performance. Further analysis shows reduced expert redundancy and the emergence of *modality-specific* and *modality-agnostic* experts, indicating implicit representation specialization within an autoregressive framework that can be leveraged for data and parameter-efficient multimodal generation. These results suggest that redundancy-aware MoE models can support data- and parameter-efficient multimodal generation, providing insight into how autoregressive objectives can serve as a strong foundation for next-generation multimodal models.

1 INTRODUCTION

The success of Large Language Models (LLMs) (Kim et al., 2024; Trinh et al., 2024) has motivated efforts to extend autoregressive next-token prediction beyond text to multimodal generation. While diffusion-based models have historically dominated image generation (Ho et al., 2020; Dhariwal & Nichol, 2021), autoregressive next-token prediction is now being successfully applied to this task (Liu et al., 2023; Sun et al., 2024; Jin et al., 2024). In parallel, alternative paradigms such as predictive encoders (Assran et al., 2023) and latent-space modeling have highlighted the importance of representation structure and capacity allocation for efficient multimodal learning.

A central challenge in autoregressive multimodal fine-tuning is that adapting a pre-trained *text-only* model to new modalities often degrades its original language generation abilities, or requires introducing substantial modality-specific parameters or additional text data (Team, 2024; He et al., 2024), rendering existing approaches both *data and parameter-inefficient*. These limitations

raise questions about whether existing models possess sufficient latent representation capacity to accommodate new modalities efficiently, and how this capacity should be utilized during continual multimodal pre-training. In this work, we address this challenge by leveraging the inherent redundancy in large language models. We posit that this redundancy can be repurposed as a source of additional latent representational capacity required to learn a new modality. Specifically, we adopt the Mixture-of-Experts (MoE) architecture (Shazeer et al., 2017), which is particularly appealing due to **(1)** the presence of substantial latent representational redundancy across experts (Chen et al., 2022; Sarkar et al., 2024; Li et al., 2024), and **(2)** the ability to allocate modality-specific representational pathways through expert routing. This design serves a dual purpose. First, repurposing redundant experts provides additional capacity for learning new modalities without introducing a large number of new parameters. Second, the routing flexibility of MoEs offers a parameter-efficient alternative to explicitly adding modality-specific modules. To further mitigate catastrophic forgetting during continual multimodal fine-tuning, we adopt *Partial* Low-Rank Adaptation (PLoRA) (Dong et al., 2024), where low-rank adapters (Hu et al., 2022) are updated exclusively using tokens from the new modality. This enforces modality-conditional adaptation within an autoregressive framework, preserving original language generation capabilities without requiring additional text-only fine-tuning.

Our method efficiently extends a pre-trained text-only MoE to multi-modal generation. This allows us to achieve strong image generation performance (see Fig. 2) using modest training data and compute budget, and without compromising the model's original language capabilities (see Tab. 1). Our experiments show our approach delivers competitive performance with only 7.5 million training samples, which is an order of magnitude lower than existing approaches (He et al., 2024). The overview of our approach is presented in Fig. 1.

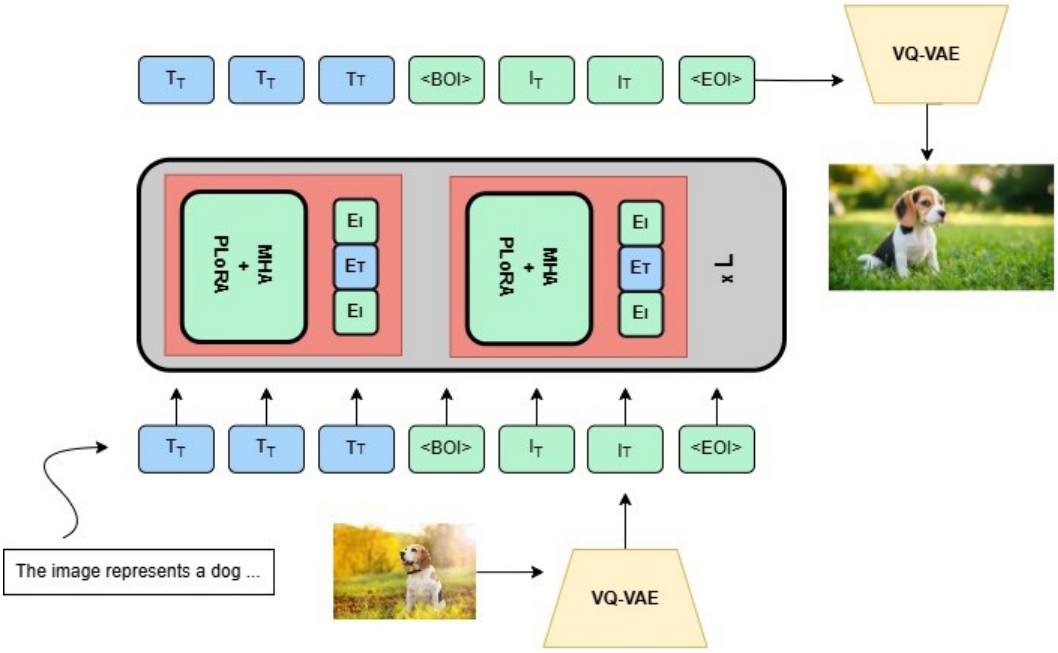


Figure 1: **Overview of the learning process:** The input image is tokenized into discrete tokens using a VQ-VAE encoder and combined with text tokens, separated by special tokens indicating the start and end of the image tokens. The LLM is trained using the next-token prediction objective, and the generated image is reconstructed using the VQ-VAE decoder. **Architecture:** The Partial LoRA (PLoRA) parameters are introduced in the attention (MHA) module, resulting in modality-specific specialization of experts.

2 Exploiting MoE Redundancy for Multimodal Generation

Preserving Language Abilities: Previous methods that extend LLMs to multiple modalities often degrade original language performance by updating weights with both image and text tokens (He

et al., 2024). We hypothesize that a pre-trained language model need only adapt to the new modality, leaving language pathways unchanged. To this end, we adopt Partial LoRA (PLoRA) (Dong et al., 2024), which applies low-rank adapters exclusively to image tokens. These adapters are introduced in the query, key, value, and output projection layers of the transformer decoder. During training, the adapters, MoE router, and experts are made trainable, enabling reallocation of representational capacity for the new modality. Formally, for an input x with image tokens x_v and text tokens x_t, text tokens are processed using the original weights W_o, while image tokens are processed using W_o augmented with trainable low-rank weights $W_B W_A$, as shown in Equation 1.

$$\begin{aligned} \hat{x}_v &= (W_o + W_B W_A) x_v + B_o, \\ \hat{x}_t &= W_o x_t + B_o, \\ \hat{x} &= [\hat{x}_v, \hat{x}_t]. \end{aligned} \tag{1}$$

We anticipate this approach will direct image tokens to previously underutilized experts and **inducing modality-specific representational pathways**, while preserving language abilities. **Note that** PLoRA is an architecture-agnostic fine-tuning procedure and can be applied to any transformer-based model, and is not exclusive to MoEs.

Continual Pre-Training with Multi-Modal Data: We perform fine-tuning on multi-modal data after introducing low-rank adapters (PLoRA), and initializing new parameters in the embedding and head layers using the Gromov-Wasserstein initialization (Mémoli, 2011). We divide our training process into two stages: **Low-Res Training** and **High-Res Training**. Low-Res training (4M 256×256 samples) focuses on establishing coarse cross-modal representational alignment, while High-Res training (3.5M 512×512 samples) refines visual fidelity and spatial detail. For both stages, we use high-quality, photorealistic images with detailed captions generated by Share-Captioner (Chen et al., 2025).

Overall, this design can be viewed as an autoregressive multimodal model in which expert routing and modality-conditional adaptation implicitly factorize representations across modalities. Unlike predictive encoders or diffusion-based approaches, this factorization emerges from capacity allocation within a standard next-token prediction objective, rather than from explicit latent forecasting or iterative denoising.

3 Experiments and Results

Experimental Settings: Our experiments are based on the LLaMA-MoE (4/16) model (Zhu et al., 2024), which is an MoE variant of LLaMA-2-7B. This model uses 16 experts per layer and activates the top four for each input token (3.5B activated params). We introduce low-rank adapters with a rank of 64 and apply rank stabilization (Kalajdzievski, 2023). We use the AdamW optimizer with a learning rate of 2e-4, decaying to 2e-5 with a cosine schedule and 1,000 warmup steps. We fine-tune in multiple stages with 7.5M image-text pairs of varying resolutions (256×256 and 512×512).

Preserving Language Representations: Results in Tab. 1 show that modality-specific routing with LLaMA-MoE + PLoRA preserves the LLM's original language capabilities. Compared to the original LLaMA-MoE and a variant using standard LoRA on both image and text tokens, naive LoRA incurs a substantial performance drop of 15.35%. In contrast, PLoRA yields a negligible degradation of only 0.14%, indicating that redundant capacity can be repurposed for multimodal learning without disrupting core language representations. These results indicate that modality-conditional adaptation preserves pre-trained language representations, allowing multimodal learning without revisiting large-scale text data, satisfying constraint C1.

Multi-Modal Fine-Tuning Reduces Expert Redundancy: Our initial hypothesis posited that inherent Mixture-of-Experts (MoE) redundancy provides the latent capacity necessary to accommodate a new modality. To quantify this, we utilize Expert Co-Activation (ECA) (Muennighoff et al., 2024) as a proxy for intra-expert redundancy. ECA is defined as the normalized proportion of instances where two experts activate simultaneously; high values indicate frequent joint activation and potential functional overlap. We expect a reduction in average ECA if redundant experts are successfully repurposed for multimodal generation.

Fig. 3 illustrates the average ECA across all experts in each layer. Firstly, we observe that experts in the original pre-trained MoE exhibit substantial redundancy across layers. For instance, **layers 21**

Figure 2: Example generated samples using our approach exhibiting high fidelity and strong textual coherence. See Appendix A.4 and A.2 for text prompts and more generated samples, respectively.

Model \ Task	Winogrande	NQ_Open	Hellaswag	MMLU	Lambada	Arc-e	Arc-c	Piqa	SciQ	Average
LLaMA-MoE	65.60	20.30	73.48	39.91	69.50	65.82	44.20	77.90	87.60	60.47
LLaMA-MoE + LoRA	43.27	10.18	59.44	22.08	52.36	54.27	32.10	67.38	65.04	45.12 (-15.35 ↓)
LLaMA-MoE + PLoRA (Ours)	65.35	20.06	73.30	39.73	69.45	65.60	44.12	77.85	87.53	60.33 (≈-0.14)

Table 1: Performance comparison on text benchmarks for the original LLaMA-MoE, LLaMA-MoE fine-tuned using LoRA, and PLoRA. While naive LoRA significantly degrades text performance, PLoRA can be seen to preserve the original text capabilities (-15.35% and -0.14% respectively as compared to the original average performance).

and 25 show high ECA values ($\approx$0.70), depicting high redundancy among the experts. In contrast, after multi-modal fine-tuning, this redundancy is markedly reduced, especially in the initial layers of the model. **Layer 0** shows extremely low ECA ($\approx$0.10). Overall, this reduction implies that the model has effectively leveraged inherent redundancy as the latent capacity to learn the new modality, adding further support for our hypothesis.

Emergence of Modality-Specific and Shared Representations: We analyze routing preferences across modalities in Fig. 4 by measuring how frequently image and text tokens are routed to each of the 16 experts across layers. Tokens from different modalities exhibit *pronounced routing exclusivity*, particularly in the early and late layers (Layers 1 and 31), where experts preferred by image tokens are rarely selected by text tokens, and vice versa. In contrast, intermediate layers (Layers 5 and 14) show substantially weaker specialization, indicating greater sharing of representational capacity. Together, these patterns demonstrate the emergence of modality-specific experts following multimodal fine-tuning, with specialization concentrated at the representational boundaries of the network.

4 Conclusion

In this work, we study how pre-trained, text-only autoregressive LLMs can be extended to multimodal generation without sacrificing their original language capabilities and without introducing substantial additional parameters. By leveraging expert redundancy in Mixture-of-Experts architectures and applying modality-conditional adaptation via Partial LoRA, we show that multimodal generative

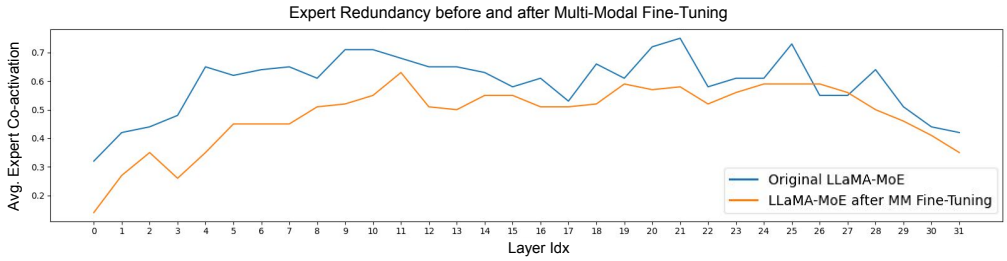


Figure 3: Average expert redundancy (co-activation) across each layer before and after multi-modal fine-tuning. The observed reduction in average expert redundancy after fine-tuning indicates that redundant experts were leveraged to learn the new modality.

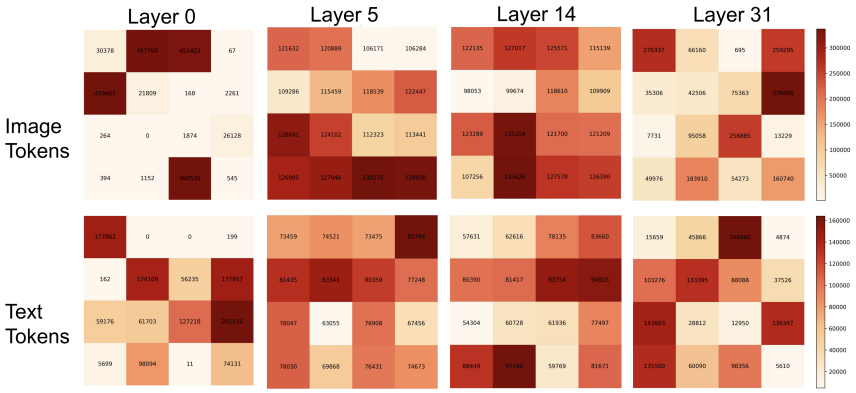


Figure 4: Number of tokens routed (routing preferences) to each of the 16 experts in the initial, middle and final layers of LLaMA-MoE.

abilities can be acquired efficiently under a standard next-token prediction objective. Our empirical results demonstrate high-fidelity text-to-image generation with negligible degradation in language performance, while detailed routing and co-activation analyses reveal reduced expert redundancy and the emergence of both modality-specific and shared experts. These findings suggest that multimodal training repurposes existing latent capacity into more structured and specialized representations, rather than relying on explicit architectural expansion.

Viewed more broadly, this work provides evidence that appropriately structured autoregressive models can implicitly specialize and factorize representations through capacity allocation mechanisms, offering some of the representational benefits often associated with alternative paradigms such as predictive encoders or latent-space modeling. **Future Directions:** As a work in progress, this study opens several directions for future research, including scaling to larger MoE models (Muennighoff et al., 2024), exploring more expressive routing mechanisms (Huang et al., 2024; Dai et al., 2024), and investigating hybrid training strategies that combine autoregressive objectives with latent forecasting or refinement-based approaches. Through scaling to larger datasets, future work can also perform a direct comparison with competing approaches (He et al., 2024). Finally, being part of an ongoing study, the current scope of evaluations are limited to a specific set of experiments. Future work would incorporate a comprehensive set of experiments including notable evaluation datasets (Ghosh et al., 2023; Hu et al., 2024) and baselines (Liu et al., 2023; Team, 2024; Deng et al., 2025).

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

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*, pp. 370–387. Springer, 2025.

Tianyu Chen, Shaohan Huang, Yuan Xie, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, and Furu Wei. Task-specific expert pruning for sparse mixture-of-experts. *arXiv preprint arXiv:2206.00277*, 2022.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Wanggui He, Siming Fu, Mushui Liu, Xierui Wang, Wenyi Xiao, Fangxun Shu, Yi Wang, Lei Zhang, Zhelun Yu, Haoyuan Li, et al. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. *arXiv preprint arXiv:2407.07614*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Quzhe Huang, Zhenwei An, Nan Zhuang, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Liwei Chen, Songfang Huang, and Yansong Feng. Harder tasks need more experts: Dynamic routing in moe models. *arXiv preprint arXiv:2403.07652*, 2024.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin CHEN, Chengru Song, dai meng, Di ZHANG, Wenwu Ou, Kun Gai, and Yadong MU. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FlvtjAB0gl.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora, 2023. URL https://arxiv.org/abs/2312.03732.

Jiyeong Kim, Kimberly G Leonte, Michael L Chen, John B Torous, Eleni Linos, Anthony Pinto, and Carolyn I Rodriguez. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digital Medicine*, 7(1):193, 2024.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient SMoe with hints from its routing policy. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=eFWG9Cy3WK.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

Soumajyoti Sarkar, Leonard Lausen, Volkan Cevher, Sheng Zha, Thomas Brox, and George Karypis. Revisiting smoe language models by evaluating inefficiencies with task specific expert pruning. *arXiv preprint arXiv:2409.01483*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

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.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. Llama-moe: Building mixture-of-experts from llama with continual pre-training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15913–15923, 2024.

A APPENDIX

A.1 MULTI-MODAL GENERATION VIA UNIFIED ARCHITECTURE

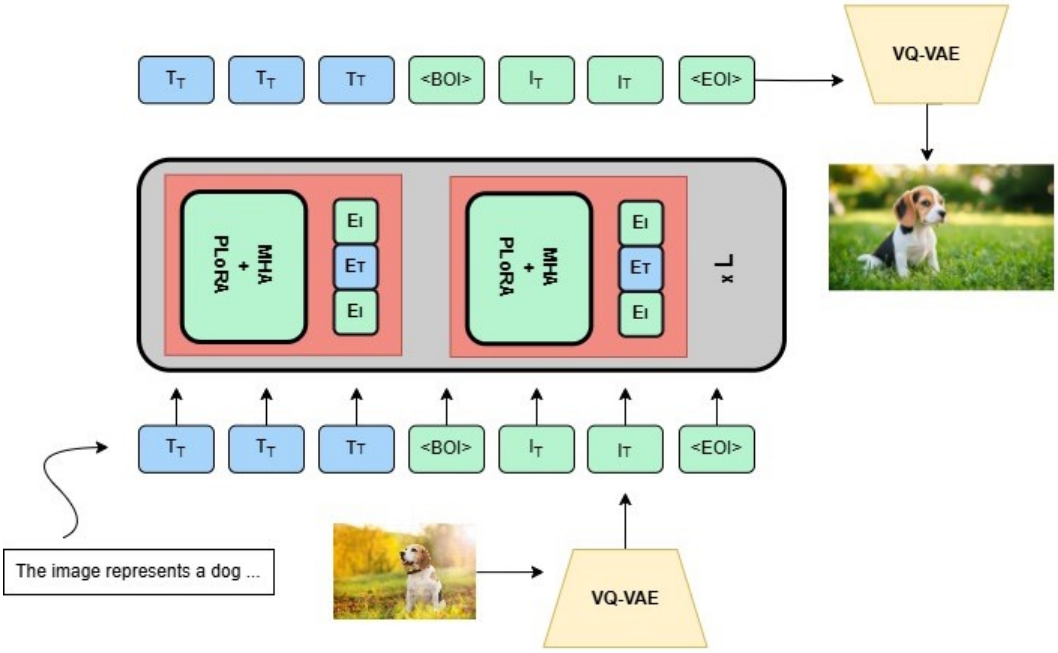
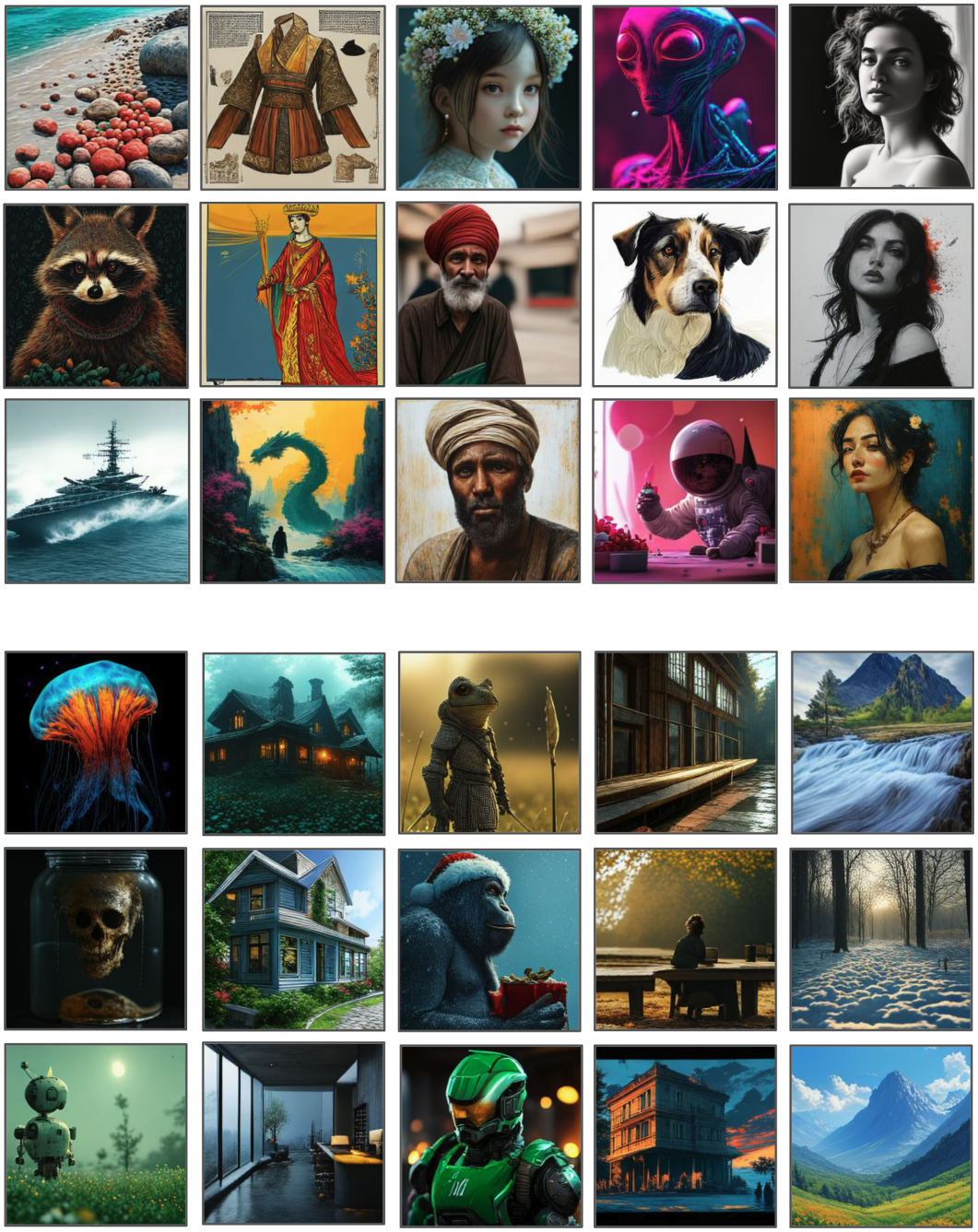


Figure 5: Overview of the learning process: The input image is tokenized into discrete tokens using a VQ-VAE encoder and combined with text tokens, separated by special tokens indicating the start and end of the image tokens. The LLM is trained using the next-token prediction objective, and the generated image is reconstructed using the VQ-VAE decoder.

To achieve a unified architecture for multiple modalities (image and text in our case), the first step is to convert the new input modality into discrete token sequences using modality-specific encoders, which can then be trained with the standard next-token prediction loss employed by LLMs. For images, this tokenization is typically performed using a Vector-Quantized Variational AutoEncoder (VQ-VAE) (Van Den Oord et al., 2017). A VQ-VAE transforms the input image pixels x into a corresponding feature map f and assigns each vector $f^{(i,j)}$ in the feature map to the index $q_{(i,j)}$ of its closest codebook vector $z^{(i,j)}$. During decoding, the indices $q_{(i,j)}$ are mapped back to their respective codebook vectors $z^{(i,j)}$ which are then reconstructed into the image pixels $\hat{x}$ by the decoder. We employed the image tokenizer in Sun et al. (2024), which has a codebook size of 16384.

In order for the LLM to interpret these *new* tokens, we proceed as follows. We expand the tokenizer's vocabulary by adding 16384 tokens corresponding to images, along with two special tokens, **<|boi|>** and **<|eoi|>**, which indicate the beginning and end of an image in the input sequence, respectively. To encode and decode these tokens, we further enlarge the embedding and head layers of the LLM. Formally, let the number of new tokens (image tokens plus special tokens) be T, let $|V_t|$ denote the size of the text vocabulary, and let d denote the embedding dimension. The original embedding and head layers have shapes $|V_t| \times d$. After incorporating the new tokens, the total parameter count becomes $(|V_t| + T) \times d$, meaning that an additional $T \times d$ parameters have been introduced. For example, in our implementation using the "*LLaMA-MoE (4/16)*" model where the embedding dimension is 4096 and the original vocabulary size is 32000, the expansion resulted in the addition of $16386 \times 4096 \approx 67M$ parameters. **Note that** this expansion is a standard procedure when incorporating a new modality in a unified architecture and **does not indicate** any parameter inefficiency in our approach.

A.2 More Generated Samples



A.3 Parameter, Data, and Compute Budget

We further contrast our approach with MARS in Tab. 2 in terms of parameter, data and compute budget.

Parameter Budget: The parameter count of the base LLM employed in both approaches is the same (7B), however, MARS employs a VLM, which already understands the image modality. The number of parameters introduced specifically for learning the new modality is denoted by "**New Modality Params**". In the case of MARS, the SemVIE module introduces 7B parameters, effectively doubling

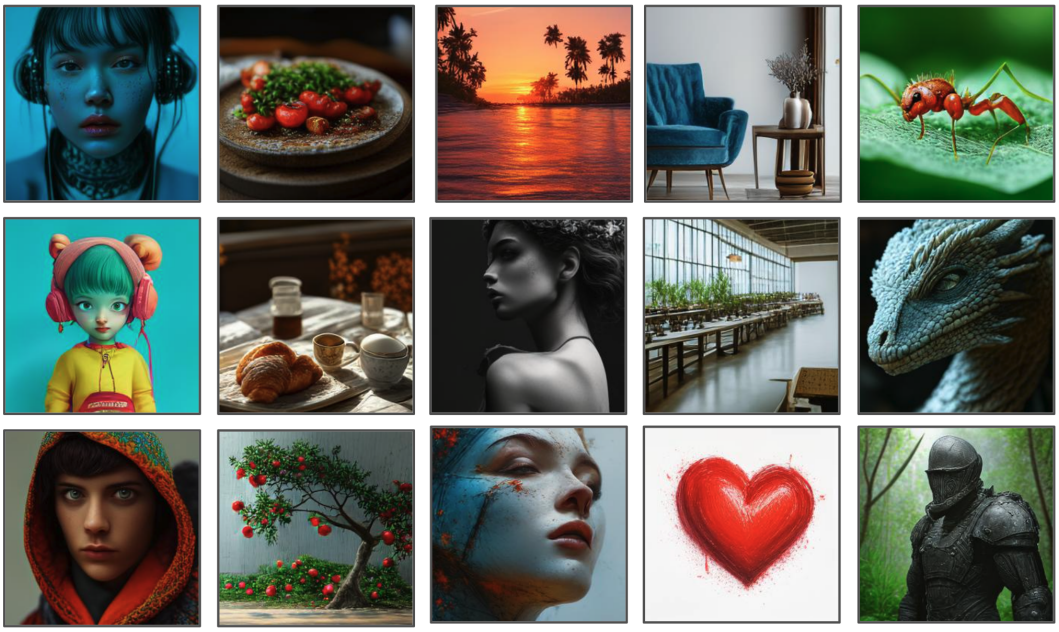


Figure 6: More examples of generated samples.

	MARS	LLaMA-MoE-PLoRA
Base LLM	Qwen-7B	LLaMA-2-7B
Base LLM Type	Multi-Modal VLM	Uni-Modal LLM
Base LLM Params	7B	7B
Activated Params	7B	3.5B
New Modality Params	7B	0.0083B
Training Data	250M	7.5M
A100 GPU Days	587	180 (approx)

Table 2: A comparison with MARS (He et al., 2024) on Parameter, Data, and Compute Budget.

the parameter count. For LLaMA-MoE, PLoRA parameters account for only 0.008B parameters (875x reduction).

Data Budget: MARS was trained on 250M samples where 200M samples were used for Stage-1 training and an additional 50M high-quality samples were employed for Stage-2 training. On the contrary, our approach employed just 7.5M high-quality samples.

Compute Budget: MARS conducted training for a total of 587 A100 GPU days as opposed to just 180 A100 GPU days for our approach. **Note that** we provide an approximation of "A100 GPU days" since our training was conducted on Nvidia L40 GPUs, which have significantly smaller VRAM (48GB as compared to 80GB in A100s).

A.4 Generation Prompts

In this section, we present the text prompts used for generating the images in Fig. 2.

Top Row | Cols 1-5

1. In the image, there is a young woman who is the main subject. She is adorned in a vibrant red dress that contrasts beautifully with the black wall behind her. The dress features a ruffled neckline and sleeves, adding a touch of elegance to her attire. On her head, she wears a wide-brimmed red hat, which matches her dress and adds a pop of color to the scene. The woman's gaze is directed off to the side, giving her a thoughtful and serious expression. This, combined with her direct gaze into the camera, creates a captivating portrait. Her hair, styled in loose waves, frames her face and complements her overall look. The image does not contain any text or other discernible objects. The focus is solely on the woman, her attire, and her expression. The relative position of the woman to the black wall suggests she is standing quite close to it. The image captures a single moment in time, with no indication of movement or action. It's a still portrait that tells a story through its subject and her attire.
2. The image captures a close-up of a woman's face, bathed in soft light. Her eyes are gently closed, and her lips are slightly parted, as if she's about to speak. The focus is on her nose and forehead, which are adorned with small droplets of water. The droplets, glistening under the light, add a sense of freshness to her appearance. The background is a dark blue-green color, providing a stark contrast to the woman's skin tone. This contrast accentuates the details of her face, making them stand out even more. The image does not contain any text or other discernible objects. The relative position of the woman's face to the background suggests that she is the main subject of this image. The overall composition of the image is simple yet striking, with the woman's face being the focal point.
3. In the image, a white and gray cat with striking blue eyes is the main subject. The cat's fur is long and shaggy, giving it a fluffy appearance. Its ears are pointed upwards, adding to its alert and curious expression. The cat is looking to the left of the frame, as if something has caught its attention. The cat is positioned in front of a window, which is blurred in the background, suggesting a depth of field effect from the camera. The window allows light to filter into the room, casting a soft glow on the cat. The cat's gaze and the direction of the light create a sense of interaction between the viewer and the scene. The image captures a quiet moment in the cat's day, providing a glimpse into its world. The colors, lighting, and composition all contribute to a serene and captivating image.
4. The image presents a close-up view of a woman's face, captured in profile. Her eyes are gently closed, and her lips are slightly parted as if she's about to speak or sing. The woman's face is adorned with a vibrant array of colors and patterns, creating a mosaic-like effect that covers most of her visage. The colors span a wide spectrum, including hues of blue, orange, red, and yellow, which stand out vividly against the stark white of her skin. The patterns on her face are intricate and varied, with geometric shapes and swirls interspersed throughout. These patterns add a dynamic element to the image, making the woman's face appear as if it's telling a story or expressing an emotion. The background of the image is a light beige color, speckled with small black dots scattered randomly across it. This backdrop provides a neutral canvas that allows the colors and patterns on the woman's face to take center stage. Overall, the image is a striking piece of art that uses color, pattern, and perspective to create a captivating visual narrative. The woman's face, with its colorful mosaic-like design, is the focal point of the image, drawing the viewer's attention and inviting them to explore the story behind the artwork.
5. The image presents a captivating digital art piece featuring a woman's face. The woman's face, which is the central focus of the image, is rendered in a realistic style. Her features are accentuated with a palette dominated by shades of blue and gray, lending an air of tranquility to her expression. Her eyes, painted in a deep shade of blue, gaze upwards and to the left, as if lost in thought or perhaps gazing at something beyond the frame of the image. Her lips, painted in a soft pink hue, add a touch of warmth to the cool color scheme. The background of the image is a stark white, providing a contrast that makes the woman's face stand out. Adding an element of intrigue to the image are black lines and splatters that surround the woman's face. These elements appear to be abstract brushstrokes, further enhancing the digital art style of the piece. Overall, the image is a beautiful blend of color and form, with each element carefully placed to create a harmonious composition. The use of color and form to convey emotion and mood is a testament to the skill and creativity of the artist.

Middle Row | Cols 1-5

1. The image presents a close-up view of a young woman's face, captured in a digital art style. Her eyes, a striking shade of blue, are the focal point of the image, radiating a sense of calm and tranquility. Her hair, a vibrant shade of pink, is styled in loose curls that frame her face, adding a touch of whimsy to the overall composition. She is adorned with a pair of silver earrings, which add a subtle sparkle to her appearance. The background is a blurred mix of pink and white hues, providing a soft contrast that allows the woman's features to stand out. The image does not contain any discernible text or additional objects. The relative position of the woman to the background suggests she is centrally located within the frame. The image does not provide any information about the woman's actions, as it appears to be a still portrait. This detailed description is based on the visible elements in the image and does not include any speculative or imaginary content.
2. In the image, a man is captured in a close-up portrait. He is adorned with a red and green knit hat, which is decorated with white pom poms at the top. The hat's vibrant colors contrast beautifully with his black beard and mustache. His gaze is directed straight at the camera, creating a sense of connection with the viewer. The background of the image is blurred, drawing focus to the man. It appears to be a room filled with Christmas lights, adding a festive atmosphere to the scene. The image does not contain any discernible text. The man's position relative to the background suggests he is standing in front of the lights. The overall composition of the image places the man as the central focus, with the Christmas lights serving as a secondary element in the background.
3. In the image, a turtle is the main subject, captured in a close-up shot. The turtle 's shell is a striking pattern of black and orange, adorned with intricate designs that add to its charm. The turtle is situated on a bed of small rocks, which provide a contrasting texture to the smoothness of the turtle's shell. The rocks are scattered around the turtle, some closer to the camera and others further away, creating a sense of depth in the image. The background is a soft blur of green foliage, providing a natural backdrop that allows the turtle to stand out. The sun is shining brightly in the top left corner of the image, casting a warm glow over the scene and creating a lens flare that adds a touch of magic to the image. Overall, the image captures a serene moment in nature, with the turtle as the star of the scene. The turtle's vibrant colors, the detailed patterns on its shell, and the tranquil setting all combine to create a captivating image
4. In the image, a young girl with curly hair and glasses is the central figure. She is dressed in a white blouse and a blue apron, adding a touch of charm to her appearance. In her hands, she holds two distinct objects - a black cat and a yellow orb. The cat, with its fur as dark as night, is comfortably perched on her shoulder, while the orb, glowing with a warm yellow light, is held in her other hand. The setting appears to be a cozy room, filled with various objects that give it a lived-in feel. A bookshelf filled with books suggests a love for literature, while a clock on the wall indicates the passage of time. A plant adds a touch of greenery to the room, creating a harmonious blend of indoor and outdoor elements. The precise locations of these objects create a well-balanced composition, with the girl and her cat at the center, drawing the viewer's attention. The image does not contain any discernible text. The relative positions of the objects suggest a quiet, peaceful moment captured in time. The girl, the cat, and the orb are all in close proximity, suggesting a bond between them. The room serves as a backdrop, framing the scene and adding depth to the image.
5. In the image, a woman is captured in a close-up shot, her face adorned with intricate makeup and a traditional Native American headdress. The headdress, a striking feature, is decorated with feathers in hues of brown, red, and white. The woman's gaze is directed towards the left side of the frame, her eyes accentuated by long, dark lashes. In her hand, she holds a pipe, a symbol often associated with Native Americans. The background of the image is blurred, drawing focus to the woman. However, it's discernible that the setting is outdoors, possibly a forest or a field, adding a natural element to the composition. The image does not contain any discernible text. The relative positions of the objects suggest a sense of depth, with the woman in the foreground and the forest or field in the background. The woman, the pipe, and the headdress are the main elements in the image, while the background provides context to the setting. The image does not provide any information that allows for a confident count of the objects in the background. Overall, the image captures a moment of stillness, with the woman in her traditional attire, the pipe in her hand, and the natural backdrop. The precise locations of the objects cannot be determined from the image alone.

The image does not contain any imaginary content; everything described can be confidently determined from the image itself.

Bottom Row | Cols 1-5

1. The image captures a close-up of a man's face, his features etched with the lines of age and experience. His eyes, a striking shade of blue, gaze directly into the camera, unflinching and intense. His nose, prominent and well-defined, stands out against the rest of his face. The skin of his face is weathered and wrinkled, a testament to the years he has lived. He is dressed in a black jacket, its dark color contrasting with the lighter tones of his face. The background is a blurred gray, a neutral backdrop that further emphasizes the man's face. The image does not contain any discernible text or other objects. The man's position relative to the camera and the background suggests he is the main subject of this image. The image does not provide any information about the man's actions, as he appears to be in a state of stillness. The image is devoid of any aesthetic descriptions, focusing solely on the man and his immediate surroundings.
2. The image portrays a young boy, adorned in regal attire, exuding an air of majesty and nobility. His gaze is directed straight at the camera, his expression serious, perhaps reflecting the solemnity of his attire. His head is crowned with a gold crown, which is intricately designed with a cross at its center, symbolizing authority and power. Complementing the crown, he wears a gold necklace around his neck, adding to his royal appearance. Over his shoulders, he drapes a large, ornate cape. The cape is richly decorated with gold embroidery, showcasing the meticulous craftsmanship involved in its creation. The fabric of the cape appears to be of a luxurious nature, enhancing the overall grandeur of the boy's attire. The background of the image features a green curtain, providing a stark contrast to the boy's golden attire. The curtain's texture and color add depth to the image, framing the boy and drawing attention to him as the focal point. Overall, the image captures a moment of quiet dignity and regal elegance, embodied by the young boy in his ornate attire. The precise positioning of the objects and the boy's direct gaze create a sense of engagement with the viewer, inviting them to appreciate the intricate details and the overall composition of the image.
3. In the image, a small dog with a coat of brown and white fur is the main subject. The dog's eyes, a striking shade of blue, are gazing directly into the camera, giving it a curious and alert expression. Adding to its charm is a pink collar around its neck, from which hangs a silver tag. The dog is not just any ordinary pet, it's a service dog, as indicated by the black harness it's wearing. The harness is equipped with a silver buckle, matching the silver tag on its collar. The setting of the image is equally captivating. The dog is standing on a road that appears to be at sunset. The sky, painted in hues of orange and yellow, suggests that the sun is setting, casting a warm glow over the scene. In the distance, you can see trees standing tall, their silhouettes adding depth to the landscape. Overall, the image beautifully captures a moment in the life of this service dog, set against the backdrop of a serene sunset.
4. The image presents a captivating scene of a fox, rendered in vibrant hues of orange and brown, with striking red accents on its face and ears. The fox is captured in profile, its gaze directed towards the right side of the image, as if gazing into the distance. It stands amidst a lush array of green foliage and flowers, adding a touch of nature's charm to the composition. The entire scene is encapsulated within a circular frame, lending a sense of completeness to the image. The art style is reminiscent of stained glass, with the fox and the surrounding flora intricately intertwined, creating a harmonious blend of colors and shapes. The image does not contain any discernible text or countable objects, and there are no explicit actions taking place. The relative positions of the objects suggest a serene coexistence, with the fox and the flora existing in harmony within their shared space. The image is a testament to the beauty of nature, captured in a moment of tranquility.
5. In the image, a figure clad in a futuristic suit of green and gray is seated in a chair. The suit is detailed with a chest plate. The figure's head is protected by a helmet equipped with a visor that exhibits a striking red and orange glow. The figure appears to be engrossed in reading a piece of paper that rests on their lap. The setting is a dimly lit room. The overall atmosphere of the image suggests a scene straight out of a science fiction narrative.