MARS: Mixture of Auto-Regressive Models for Fine-grained Text-to-image Synthesis

Wanggui He[1*], Siming Fu[1*], Mushui Liu[2*], Xierui Wang[2], Wenyi Xiao[2], Fangxun Shu[1], Yi Wang[2], Lei Zhang[2], Zhelun Yu[3], Haoyuan Li[1], Ziwei Huang[2], Leilei Gan[2†], Hao Jiang [1†]

[1]Alibaba Group
[2]Zhejiang University
[3]Fudan University
wanggui.hwg@taobao.com, {fusiming, lms, sherrywang, wenyixiao}@zju.edu.cn, shufangxun.sfx@alibaba-inc.com, y_w@zju.edu.cn, shufangxun.sfx@alibaba-inc.com, zhelunyu13@fudan.edu.cn, {lihaoyuan, 22351096, leileigan}@zju.edu.cn, aoshu.jh@taobao.com

Abstract

Auto-regressive models have made significant progress in the realm of text-to-image synthesis, yet devising an appropriate model architecture and training strategy to achieve a satisfactory level remains an important avenue of exploration. In this work, we introduce **MARS**, a novel framework for T2I generation that incorporates a specially designed Semantic Vision-Language Integration Expert (SemVIE). This innovative component integrates pre-trained LLMs by independently processing linguistic and visual information—freezing the textual component while fine-tuning the visual component. This methodology preserves the NLP capabilities of LLMs while imbuing them with exceptional visual understanding. Building upon the powerful base of the pre-trained Qwen-7B, MARS stands out with its bilingual generative capabilities corresponding to both English and Chinese language prompts and the capacity for joint image and text generation. The flexibility of this framework lends itself to migration towards **any-to-any** task adaptability. Furthermore, MARS employs a multi-stage training strategy that first establishes robust image-text alignment through complementary bidirectional tasks and subsequently concentrates on refining the T2I generation process, significantly augmenting text-image synchrony and the granularity of image details. Notably, MARS requires only **9%** of the GPU days needed by SD1.5, yet it achieves remarkable results across a variety of benchmarks, illustrating the training efficiency and the potential for swift deployment in various applications.

Introduction

Pre-trained large language models (LLMs) (Zhang et al. 2022; Brown et al. 2020; Wei et al. 2022; Touvron et al. 2023; Wang et al. 2023) have broadened their generative capabilities to encompass the visual domain. This advancement entails transforming pixel data into discrete tokens through a visual tokenizer, analogous to the processing of textual information, thereby integrating these tokens into the model's transformer (Vaswani et al. 2017) architecture for generative tasks. Unlike current diffusion models (Rombach et al. 2022; Podell et al. 2023a; Esser et al. 2024; Chen et al. 2023; Ma et al. 2024c,b; Liu et al. 2025), LLMs (Chang et al. 2023; Yu et al. 2023; Ding et al. 2022; Ma et al. 2024a) uniquely utilize a discrete latent space of visual tokens, crucial for merging visual and linguistic modalities.

Auto-regressive models for text-to-image generation models,*e.g.*, Parti (Yu et al. 2023), CogView2 (Ding et al. 2022), and Unified-io2 (Lu et al. 2024) have extended their generative scope to encompass the visual domain, facilitating the creation of images. These models integrate pre-trained LLMs within a unified architecture, enabling the simultaneous interpretation of both linguistic and visual inputs. akin to challenges encountered in previous investigations (Lu et al. 2024; Yu et al. 2023), a significant impediment arises from the inherent distributional bias of LLMs, which are predominantly trained on textual data, potentially leading to a pronounced distributional shift when adapting to text-image pair datasets. This shift has the potential to provoke catastrophic forgetting, consequently impairing the LLMs' primary competency in text generation tasks. *The aforementioned discourse prompts a pivotal inquiry: Is it feasible to preserve the natural language processing proficiency of LLM while concurrently endowing it with state-of-the-art visual comprehension and generation capabilities?*

In response to this challenge, **different from** the popular approaches (Yu et al. 2023; Lu et al. 2024) for parameter sharing across multiple modalities, we present MARS, an innovative framework predicated on an auto-regressive model architecture akin to that of pre-trained LLMs for text-to-image synthesis. Specifically, we design the Semantic Vision-Language Integration Expert (SemVIE) module as the centerpiece of MARS to seamlessly facilitate the frozen pre-trained LLM with the trainable visual expert, thereby endowing them with exceptional visual understanding and preserving the NLP capability of pre-trained LLMs. **Therefore**, SemVIE can facilitate a comprehensive and incremental interplay between the textual and visual modalities across every layer of the model, fostering deep inte-

*These authors contributed equally.
†Corresponding Authors.

Figure 1: The generated samples from MARS display extraordinary quality, marked by an impressive degree of fidelity and precision in their adherence to the provided textual descriptions.

gration that yields images closely aligned with their textual descriptors. In addition, to endow the model with exceptional bilingual (Chinese and English) command adherence and to enhance the foundational quality of image resolution, we have formulated a multi-stage training strategy. This approach commences with the generation of low-resolution images and progressively advances toward the creation of high-resolution images with intricately aligned text. **The third phase** of our design—the High-Resolution Refinement stage—judiciously employs the Next-K Token Prediction method to augment the stability of image generation and to improve computational efficiency.

Through rigorous training on paired image-text datasets, MARS augments the generative capabilities of LLMs to include sophisticated text-to-image translations. As demonstrated in Fig. 1, MARS exhibits a pronounced ability to generate images with intricate visual details, such as animal fur, plant foliage, and facial features, underscoring its potent text-to-image generation proficiency. Remarkably, with a mere 587 A100 GPU days, equating to only **9%** of the training duration required by Stable Diffusion v1.5, MARS demonstrates its superiority over existing large-scale text-to-image (T2I) models. Overall, our contributions can be encapsulated as follows:

- We present MARS, an innovative framework adapted from auto-regressive pre-trained LLMs for T2I generation tasks. To ensure the preservation of NLP capacities while also equipping the model with advanced visual generation and comprehension abilities, we design a module named SemVIE, which adds parallel visual experts to the attention blocks of pre-trained LLM. Therefore, MARS amplifies the flexibility of autoregressive methods for T2I generation and joint image-text synthesis, with the potential expansibility to **any-to-any** tasks.
- We propose a multi-stage refinement training strategy that significantly enhances MARS' robust instruction-following capability and its ability to generate high-quality images with rich details. Crucially, the High-Resolution Refinement stage substantially improves not only the visual fidelity of the generated images but also the efficiency of the inferential process.
- MARS shows great ability in prompt understanding and following, *e.g.* long and complex nature language inputs. Moreover, it possesses the **bilingual** capacity to follow prompts in both English and Chinese. The framework's performance is verified across an array of evaluative measures, *i.e.* MS-COCO benchmark, T2I-CompBench, and Human Evaluation.

Related Works

Text-to-Image Generation Models. Text-to-image generation aims to create images based on given textual descriptions. Recent diffusion-based models (Ho, Jain, and Abbeel 2020; Song et al. 2020) have demonstrated exceptional performance in image generation, offering improved stability and controllability. These models operate by introducing Gaussian noise to input images in a forward process and subsequently generate high-quality images with intricate details and diversity through an inverse process starting from random Gaussian noise. Models like GLIDE (Nichol et al. 2022) and Imagen (Saharia et al. 2022) utilize the CLIP (Radford et al. 2021) text encoder to enhance image-text alignment. Latent Diffusion Models (LDMs) (Rombach

et al. 2022) have been proposed to shift the diffusion process from pixel space to latent space, thereby enhancing efficiency and image quality. Furthermore, recent advancements such as SD-XL (Podell et al. 2023a), and DALL-E 3 (Betker et al. 2023) have significantly improved image quality and text-image alignment by employing various approaches, including innovative training strategies and scaling of training data. Furthermore, with the diffusion model framework transitioning from a U-Net structure towards a transformer-based architecture DiT (Peebles and Xie 2023). PixArt-α (Chen et al. 2023), SD-3.0 (Esser et al. 2024), and Lumina-T2X (Gao et al. 2024) achieve exceptional performance through the integration of DiT.

Auto-regressive Model for Visual Generation. Auto-regressive Models (Brown et al. 2020; Touvron et al. 2023) have been adeptly repurposed for the synthesis of visual media, including images (Chang et al. 2023; Ding et al. 2022) and videos (Zhang, Li, and Bing 2023). The process begins with a visual tokenizer function implemented by VQ-VAE (Van Den Oord, Vinyals et al. 2017) or VQ-GAN (Esser, Rombach, and Ommer 2021), f, which effectively converts visual stimuli into a sequence of discrete tokens. Subsequently, X is linearized into a one-dimensional token sequence via raster scan order, which is then introduced to a language-model transformer to facilitate generative modeling. Current auto-regressive models include notable architectures such as ImageGPT (Chen et al. 2020), DALL-E (Ramesh et al. 2021; Rombach et al. 2022), and Parti (Yu et al. 2022). AR model anticipates the subsequent token based on a sequence of antecedent tokens, supplemented by additional conditional data c, and adheres to a categorical distribution for $p_\theta(x_i|x_{<i}; c)$. **However**, the aforementioned methods employ shared parameters across various modalities, which can result in domain shift that may subsequently lead to a decline in natural language processing performance, as well as the occurrence of the logit shift phenomenon. In this paper, we propose a solution predicated on auto-regressive generation, underpinned by a distinctive model architecture design, to enhance quality and facilitate interactive text-guided synthesis.

Methodology

Textual and Image Discrete Tokenization

In this paper, Qwen-7B (Bai et al. 2023), a pre-trained LLM, serves as the foundational linguistic framework, tokenizing textual data into representative tokens, r_t. Simultaneously, an encoder inspired by the VQ-GAN architecture (Lee et al. 2022) transforms the image $x \in \mathbb{R}^{3\times H\times W}$ into a feature map $f_v \in \mathbb{R}^{K\times D}$, where $K = H \times W/P^2$ with P set to 16, and D representing the feature dimension. This feature map is then quantized using the VQ-GAN codebook, mapping it to a sequence of 256 tokens, each representing a 16×16 pixel segment. The visual codebook comprises 8,192 unique codes, denoted as r_v.

In MARS, these visual tokens are integrated with textual tokens, forming a multimodal vocabulary. The original LLM vocabulary contains 151,936 entries, and with the addition of the visual codebook and 6 special tokens, this expands to 160,136 entries. Visual tokens in MARS are treated equally to textual tokens, with their initial embeddings derived from the mean embedding of pre-trained textual tokens, providing a base for cross-modality integration.

Semantic Vision-language Integration Expert

The MARS architecture incorporates L layers of SemVIE, which is a specialized multi-modal Mixture of Experts (mm-MoE) designed to adeptly handle both visual and semantic tokens. Central to the SemVIE are the Attention-MoE and Feed-Forward Network (FFN)-MoE modules. A dedicated routing module is strategically situated following each layer normalization step within the transformer modules. This routing mechanism is designed to allocate each input token to the corresponding expert model best equipped for its processing. A noteworthy aspect of the shared architectural framework is the universal application of the causal multi-head attention and layer normalization modules across both language and vision modalities, epitomizing a unified methodological approach to the concurrent processing of multi-modalities data. The process of Attention-MoE follows:

$$
\begin{aligned}
\hat{r_t}, \hat{r_v} &= \text{Router}(\text{LN}(\text{Concat}(r_t, r_v))) \\
\hat{r_t}^q, \hat{r_t}^k, \hat{r_t}^v &= W_Q^t(\hat{r_t}), W_K^t(\hat{r_t}), W_V^t(\hat{r_t}) \\
\hat{r_v}^q, \hat{r_v}^k, \hat{r_v}^v &= W_Q^v(\hat{r_v}), W_K^v(\hat{r_v}), W_V^v(\hat{r_v}) \\
\hat{r_q}, \hat{r_k}, \hat{r_v} &= \mathbf{C}(\hat{r_t}^q, \hat{r_v}^q), \mathbf{C}(\hat{r_t}^k, \hat{r_v}^k), \mathbf{C}(\hat{r_t}^v, \hat{r_v}^v) \\
\hat{r} &= \text{CausalAttention}(\hat{r_q}, \hat{r_k}, \hat{r_v}) + r
\end{aligned}
\tag{1}
$$

where $\mathbf{C}$ indicates concat operation, W_Q^t, W_K^t, and W_V^t are frozen and loaded from pre-trained LLM. W_Q^v, W_K^v, and W_V^v are trainable and initialized with the pre-trained semantic LLM. Then the MoE-FFN module further processes the multi-modal tokens:

$$
\begin{aligned}
\hat{r_t}, \hat{r_v} &= \text{Router}(\text{LN}(\mathbf{C}(r_t, r_v))) \\
\hat{r_t} &= \text{FFN}^t(r_t), \hat{r_v} = \text{FFN}^v(r_v) \\
\hat{r} &= \mathbf{C}(\hat{r_t}, \hat{r_v})
\end{aligned}
\tag{2}
$$

where $\mathbf{C}$ indicates concat operation, FFN^t and FFN^v share the same architecture, and FFN^v is trainable. The SemVIE module, a cornerstone of the MARS, benefits from a synergistic integration of Attention-MoE and FFN-MoE modules, enabling the effective fusion of multimodal data streams. This integration capitalizes on the profound linguistic insights afforded by the pre-trained LLM, thus leveraging the advanced language comprehension capabilities to enrich visual understanding. To enable the model to simultaneously predict visual tokens and text tokens, in addition to using the original LLM model head (referred to as the text head), we added a vision head to the model. **Notably,** the text token and the visual token are processed through the text head and vision head to obtain the logits, denoted as l_t and l_v, respectively. The logits are then concatenated along the last dimension and passed through a softmax layer to obtain the probability distribution over the vocabulary for each token.

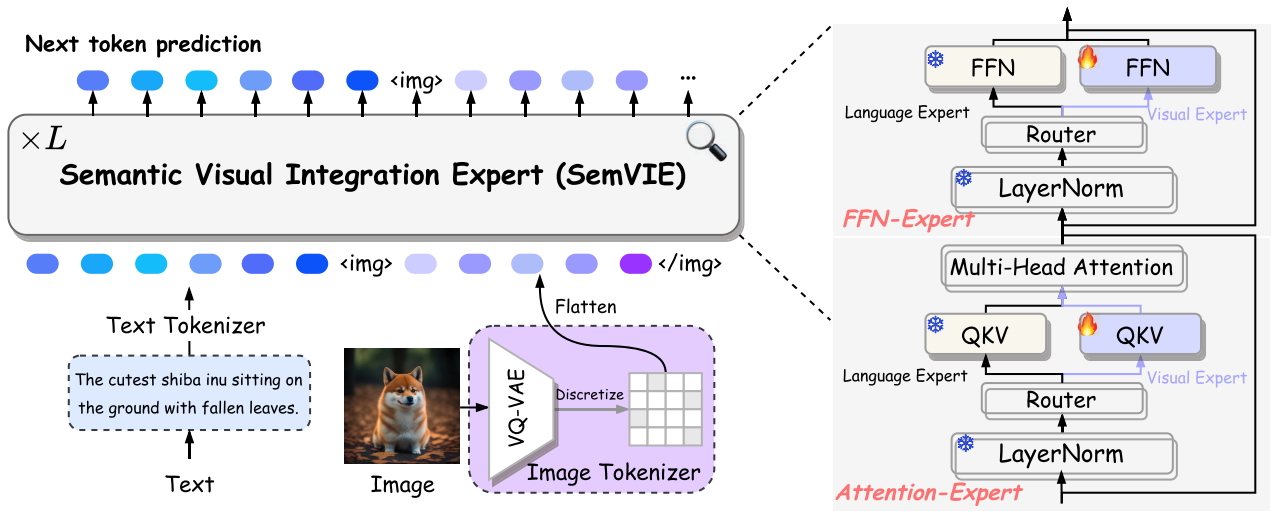

Figure 2: Overall training framework of the proposed MARS, which consists of the SemVIE modules facilitating T2I within a unified framework. An image-text pair is processed and tokenized by VQ-GAN (Esser, Rombach, and Ommer 2021) into 'vision words', which are then integrated with text tokens for joint processing in the SemVIE. The right part illustrates the multi-modal integration block, highlighting the synergistic processing of image and text data within the SemVIE, critical for the T2I task.

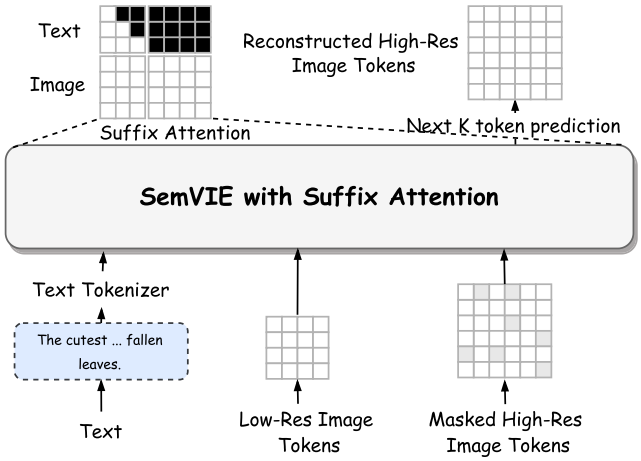

Figure 3: High-Resolution Refinement Framework: The model predicts masked tokens at higher resolution using text and low-resolution image inputs. It employs a suffix attention mechanism, enabling Next K Token Prediction (NKTP) for faster image generation.

Multi-Stage Refinement Training Strategy

Stage-I: Pre-training for Text-to-Image Alignment. We first optimize MARS by two distinct tasks: text-to-image generation and image captioning. This refinement process utilizes an auto-regressive approach for NTP. The procedure involves an extensive dataset of approximately 200 million text-image pairs, with each image conforming to a resolution of 256×256 pixels.

Stage-II: High-Quality Data Alignment. To advance the fidelity of image synthesis, this stage persists in employing an NTP for the generation of images from textual descriptions. Diverging from Stage-I, the dataset enlisted for this stage comprises 50 million pairs of text and corresponding images, each pair meticulously curated through the application of an aesthetic valuation model. The descriptive captions paired with these images originate from CogVLM (Wang et al. 2023), formulated in response to explicit directives. To mitigate potential discrepancies arising between the visual content and its textual descriptors, owing to image cropping, a standardized procedure is implemented wherein the minor axis of every image is resized to 256 pixels. This measure, taken whilst conserving the original aspect ratio, ensures the retention of comprehensive image content. However, this results in variable sequence lengths for the images. To address this, we include resolution information in the caption to specify the desired sequence lengths of the generated images.

Stage-III: High-Resolution Refinement. Inspired by the approaches of SD-XL (Podell et al. 2023b), we utilize a cascading super-resolution strategy to further enhance MARS. As illustrated in Fig. 3, the super-res model takes text and low-resolution image tokens as input and learns to predict masked tokens at a higher resolution. We use the same model architecture as in the previous Stage, changing the causal attention= to the proposed suffix attention, which applies causal attention to text tokens and bi-directional attention to image tokens. This paradigm enables the model to concurrently predict multiple image tokens within a single step, a process termed Next K Token Prediction (NKTP). This approach notably accelerates the pace of image generation. The output images have a long side of **1024 pixels** while maintaining the original aspect ratio. To control the resolution of the generated images, we apply the same strategy as in the Stage-II. Ten million triplet (low-resolution image, caption, high-resolution image) samples were used to train

Model	Attribute Binding			Object Relationship		Complex↑
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	
SD1.5 (Rombach et al. 2022)	37.65	35.76	41.56	12.46	30.79	30.80
SDXL (Podell et al. 2023a)	63.69	54.08	56.37	20.32	31.10	40.91
Composable Diffusion (Liu et al. 2022)	40.63	32.99	36.45	8.00	29.80	28.98
Structured Diffusion (Feng et al. 2022)	49.90	42.18	49.00	13.86	31.11	33.55
Attn-Exct v2 (Chefer et al. 2023)	64.00	45.17	59.63	14.55	31.09	34.01
GORS (Huang et al. 2023)	66.03	47.85	62.87	18.15	31.93	33.28
DALL-E 2 (Ramesh et al. 2022)	57.50	54.64	63.74	12.83	30.43	36.96
PixArt-α (Chen et al. 2023)	<u>68.86</u>	**55.82**	<u>70.44</u>	**20.82**	<u>31.79</u>	**41.17**
MARS (Ours)	**69.13**	<u>54.31</u>	**71.23**	<u>19.24</u>	**32.10**	<u>40.49</u>

Table 1: Evaluation results (%) on T2I-CompBench (Huang et al. 2023). The higher is better, and the best results are highlighted in **bold**.

the cascaded super-resolution model.

Experiment

Experiment Details

We employ AdamW (Loshchilov and Hutter 2017) as the optimizer, with a beta parameter of 0.95 and weight decay set at 0.1. The peak learning rate is established at 1e-4, and a warm-up strategy is employed with a ratio of 0.01. For images with a resolution of 256×256 pixels, the batch size per GPU is set at 64, while for 512×512 pixel images, it is set at 24, leading to total batch sizes of 4096 and 1536, respectively. The training utilized DeepSpeed's ZeRO-3 (Rajbhandari et al. 2020) optimization. The training epochs for Stage-I, Stage-II, and Stage-III of the model are configured to 1, 2, and 1 epochs, respectively.

Performance Comparisons and Analysis

Evaluation Benchmarks. We select three benchmarks for comparison, including MSCOCO dataset (Lin et al. 2014), T2I-CompBench (Huang et al. 2023).

MSCOCO Benchmark. We use the Frechet Inception Distance (FID) to evaluate the quality of synthesized images. As shown in 2, our proposed MARS, with only 7B trainable parameters, scores 6.92 on FID, which is a notable achievement. Compared to the auto-regressive counterpart Parti, we use fewer parameters (14B vs 20B) and smaller data sizes (0.2B vs 4.8B), achieving competitive performance (6.92 vs 7.22). Against the diffusion model SDv1.5, we achieve superior performance (6.92 vs 9.22) with less training budget (587 vs 6250 A100 GPU Days). These results highlight the efficiency of our mixture of auto-regressive models. Moreover, we utilize CLIP-Score to evaluate the alignment of textual conditions and corresponding generated images. MARS achieves 33.10 CLIPScore and 3.51 FID when the generated images are picked with the highest CLIP score, signaling its remarkable effectiveness in generating visually compelling imagery that closely adheres to the semantic content of the text prompts.

T2I CompBench Performance. The empirical data presented in Tab. 1 delineates the superior performance of our proposed MARS within the T2I-CompBench benchmark,

Method	**Arch.**	**FID-30K ↓**	**CLIP ↑**
GLIDE (Nichol et al. 2022)	Diff	12.24	-
Imagen (Ho et al. 2022)	Diff	7.27	-
SDv1.0 (Rombach et al. 2022)	Diff	-	30.50
SDv1.5 (Rombach et al. 2022)	Diff	9.22	-
MUSE (Chang et al. 2023)	Non-AR	7.88	<u>32.00</u>
DALL-E 2 (Ramesh et al. 2022)	Diff	10.39	31.40
PixArt-α (Chen et al. 2023)	Diff	7.32	-
DALL-E (Ramesh et al. 2021)	AR	28.00	-
CogView (Ding et al. 2021)	AR	27.10	-
Make-A-Scene (Gafni et al. 2022)	AR	11.84	-
Parti (Yu et al. 2022)	AR	7.23	-
Emu (Dai et al. 2023)	AR	11.70	-
CM3Leon* (Yu et al. 2023)	AR	<u>4.88</u>	-
LAVIT (Jin et al. 2023)	AR	7.40	-
UIO-2_{XXL} (Lu et al. 2024)	AR	13.39	-
MARS (Ours)	AR	6.92	32.33
MARS* (Ours)	AR	**3.51**	**33.10**

Table 2: Quantitative evaluation of MS-COCO benchmark. **Diff** means diffusion model, **AR** means auto-regressive model. The results are all from the public literature. * denotes that the results are picked from the different generated images with the best CLIP score.

underscoring its proficiency in attribute binding, delineation of object relationships, and the synthesis of intricate compositions. Notably, MARS demonstrate a marked amelioration in the fidelity of color and texture representation, achieving enhancements of +11.63% in color fidelity and +7.49% in texture accuracy relative to DALL-E 2. It further exhibited substantial advancements in spatial and non-spatial metrics compared to DALL-E 2, with improvements quantified at +6.41% and +1.67%, respectively. Moreover, when juxtaposed with the recent PixArt-α model, which integrates a T5-XL text encoder, MARS outperforms it in various dimensions. Specifically, MARS achieved the highest scores in color (69.13%) and texture (71.23%) accuracy, outperforming PixArt-α which scored 68.86% and 70.44% respectively. These results demonstrate that the incorporation of LLM representations and visual tokens within an auto-regressive framework can markedly improve the quality of generated images, as well as the alignment between the vi-

Figure 4: MARS excels in generating realistic images across resolutions and scenarios, showcasing bilingual support by effectively interpreting Chinese instructions.

Method	FID-30K ↓	CLIP ↑
w/o Visual Expert	10.13	30.14
w Visual Expert	8.24	31.03
Stage-I	8.24	31.03
Stage-II	7.02	32.21
Stage-III	6.92	32.33

Table 3: Effectiveness of SemVIE and multi-stage refinement training strategy.

sual content and its corresponding textual narratives.

Visual Analysis

Fig. 4 showcases the advanced image synthesis capabilities of the MARS framework, generating visuals with exceptional detail and fidelity to text. This success stems from sophisticated textual representations from Large Language Models (LLMs) and a structured multi-tiered training strategy that enhances the model's precision in aligning text and image. MARS's multi-stage training incrementally refines the link between textual prompts and visual outputs, producing images that reflect text intent and exhibit photorealistic detail. By leveraging LLMs' deep semantic understanding, MARS skillfully translates complex textual descriptions into coherent visual narratives, exemplifying the integration of technical efficiency and artistic expression.

Central to our language model is the Qwen architecture, designed for multilingual support and using a comprehensive dataset that includes Chinese and English. During training, we intentionally included a small yet significant amount of in-house Chinese data. As shown in Fig. 4, our model excels in Chinese text-to-image synthesis, despite the limited Chinese corpus. This indicates that MARS effectively interprets concepts across languages, ensuring that images and text merge within a unified representation space, aided by our innovative mixture mechanism.

Ablation Study

A Closer Look at SemVIE During Stage-I training, we aimed to optimize the alignment of visual and linguistic modalities by employing both text-to-image (text2image) and image-to-text (image2text) pre-training tasks. However, the shared parameter design led to the "logit drift problem," as described in Chameleon or Unified-IO-2. This issue arises from the intrinsic disparities between modalities and is evi-

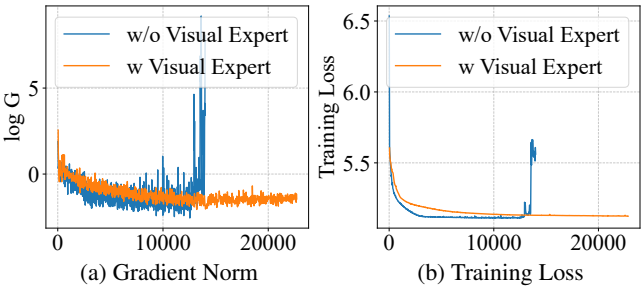

Figure 5: (a) Gradient norms and (b) Training loss on different modality mixtures. Results are obtained before and after applying the proposed SemVIE.

Figure 6: Qualitative comparison with different stages. The first row is from Stage II of MARS while the second row is from Stage III.

denced by detrimental outcomes, including a 1.89 reduction in FID, as shown in Tab. 3. To mitigate this, we introduced a specialized Visual Expert to ensure that text and visual modalities do not share parameters. This approach effectively alleviated the problem, as observed in our training results, as shown in Fig. 5. The introduction of the Visual Expert underscores the necessity for specialized architectures adept at managing the inherent challenges of multi-modal data integration.

Study on Effectiveness of Multi-Stage Refine Strategy We further explore the effect of training stages in Tab. 3. The results indicate that training in each stage positively impacts the model. On the MS-COCO benchmark, Stage II improved the FID by 1.22 compared to Stage I, and Stage III further enhanced it by 0.10 relative to Stage II. The visualizations of different stages are shown in Fig. 6. We observed that images generated during Stage I and Stage II lack detail, the images from Stage III exhibit the best quality.

Further Analysis

Image and Text Joint Generation Capability. Our proposed MARS supports simultaneous text and image generation, ensuring relevance and coherence between modalities. By maintaining the integrity of the LLM during pre-training, MARS excels in concurrent text-image tasks. For instance, in recipe generation, our pre-trained model, fine-tuned with 10,000 recipes, creates comprehensive tutorials with step-by-step instructions and corresponding illustrations. As illustrated in Fig. 7, the model generates detailed textual content and visual representations for each stage based on the recipe title and captions. MARS's seamless integration of

Figure 7: Example of Multimodal Recipe Generation from MARS. MARS can generate text and images simultaneously. The examples above demonstrate that, given a recipe title and accompanying image captions, MARS outputs the recipe steps and corresponding images end-to-end.

text and imagery can be applied to various domains requiring joint text and image generation.

Conclusion

This study introduces MARS, an innovative auto-regressive framework that not only retains the capabilities of pre-trained Large Language Models (LLMs) but also incorporates top-tier text-to-image (T2I) generation proficiency. MARS has been trained to exhibit exemplary performance in T2I tasks. We introduce the Semantic Vision-Language Integration Expert (SemVIE) module, which stands as the linchpin of MARS, streamlining the fusion of textual and visual token spaces and bringing a new insight into multi-modal learning. MARS has demonstrated superior performance in multiple benchmark assessments, such as the MS-COCO benchmark, T2I-CompBench, and human evaluations. The pre-trained Qwen model equips MARS with the ability to generate bilingual images, blending Chinese and English seamlessly. Moreover, MARS adeptly handles joint image-text generation tasks, indicating its potential for any-to-any paradigm applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62441605), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Alibaba Group through Alibaba Innovative Research Program.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM TOG*, 42(4): 1–10.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Dhariwal, P.; Luan, D.; and Sutskever, I. 2020. Generative Pretraining from Pixels. In *ICML*.

Dai, X.; Hou, J.; Ma, C.-Y.; Tsai, S.; Wang, J.; Wang, R.; Zhang, P.; Vandenhende, S.; Wang, X.; Dubey, A.; et al. 2023. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. CogView: Mastering text-to-image generation via transformers. In *NeurIPS*.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *NeurIPS*, 35: 16890–16902.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *CVPR*, 12873–12883.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2022. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *ICLR*.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 89–106.

Gao, P.; Zhuo, L.; Lin, Z.; Liu, C.; Chen, J.; Du, R.; Xie, E.; Luo, X.; Qiu, L.; Zhang, Y.; et al. 2024. Lumina-T2X: Transforming Text into Any Modality, Resolution, and Duration via Flow-based Large Diffusion Transformers. *arXiv preprint arXiv:2405.05945*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. In *ICCV*.

Jin, Y.; Xu, K.; Xu, K.; Chen, L.; Liao, C.; Tan, J.; Mu, Y.; et al. 2023. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. *arXiv preprint arXiv:2309.04669*.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive image generation using residual quantization. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, M.; Ma, Y.; Zhen, Y.; Dan, J.; Yu, Y.; Zhao, Z.; Hu, Z.; Liu, B.; and Fan, C. 2025. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *AAAI*.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *ECCV*, 423–439.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, J.; Clark, C.; Lee, S.; Zhang, Z.; Khosla, S.; Marten, R.; Hoiem, D.; and Kembhavi, A. 2024. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action. In *CVPR*, 26439–26455.

Ma, X.; Zhou, M.; Liang, T.; Bai, Y.; Zhao, T.; Chen, H.; and Jin, Y. 2024a. STAR: Scale-wise Text-to-image generation via Auto-Regressive representations. *arXiv preprint arXiv:2406.10797*.

Ma, Y.; Xu, W.; Tang, J.; Jin, Q.; Zhang, R.; Zhao, Z.; Fan, C.; and Hu, Z. 2024b. Character-Adapter: Prompt-Guided Region Control for High-Fidelity Character Customization. *arXiv preprint arXiv:2406.16537*.

Ma, Y.; Xu, W.; Zhao, C.; Sun, K.; Jin, Q.; Zhao, Z.; Fan, C.; and Hu, Z. 2024c. Storynizor: Consistent Story Generation via Inter-Frame Synchronized and Shuffled ID Injection. *arXiv preprint arXiv:2409.19624*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023a. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023b. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. arXiv:1910.02054.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *NeurIPS*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022. Finetuned language models are zero-shot learners. *ICLR*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Yu, L.; Shi, B.; Pasunuru, R.; Muller, B.; Golovneva, O.; Wang, T.; Babu, A.; Tang, B.; Karrer, B.; Sheynin, S.; et al. 2023. Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning. *arXiv:2309.02591*.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.