ByteDance | Seed

MMCORE: MultiModal COnnection with Representation Aligned Latent Embeddings

Zijie Li*, **Yichun Shi***, **Jingxiang Sun***, **Ye Wang**, **Yixuan Huang**, **Zhiyao Guo**, **Xiaochen Lian**, **Peihao Zhu**, **Yu Tian**, **Zhonghua Zhai**†, **Peng Wang***†

ByteDance Seed

*Equal Contribution, †Project Lead

Abstract

We present MMCORE, a unified framework designed for multimodal image generation and editing. MMCORE leverages a pre-trained Vision-Language Model (VLM) to predict semantic visual embeddings via learnable query tokens, which subsequently serve as conditioning signals for a diffusion model. This streamlined design effectively transfers the rich understanding and reasoning capabilities of VLMs into the visual generation process. By obviating the need for deep fusion between autoregressive and diffusion models or training from scratch, MMCORE significantly reduces computational overhead while maintaining high-fidelity synthesis.

MMCORE seamlessly integrates text-to-image synthesis with interleaved image generation, demonstrating robust multimodal comprehension in complex scenarios such as spatial reasoning and visual grounding. Comprehensive evaluations indicate that MMCORE consistently outperforms state-of-the-art baselines across a broad spectrum of text-to-image and single/multi-image editing benchmarks.

Date: April 21, 2026

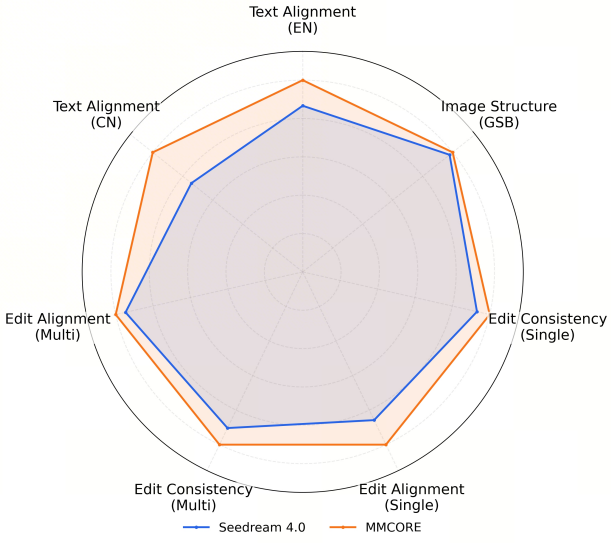
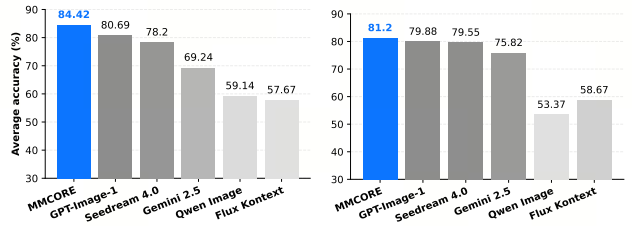
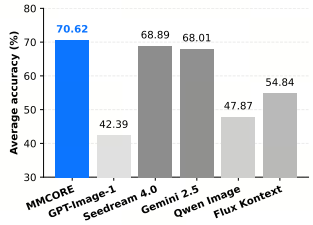


Figure 1 Human evaluation over seven metrics.



(a) Text-to-image alignment. **(b)** Image-editing alignment.



(c) Image-editing consistency.

Figure 2 DreamBench AutoEval results of various models.


MMCore
凍檸茶
SeedEdit
MMCORE
实现饮水自由，大容量800ml，一起干杯畅饮！
NOTBAD
实现饮水自由，大容量800ml，一起干杯畅饮！

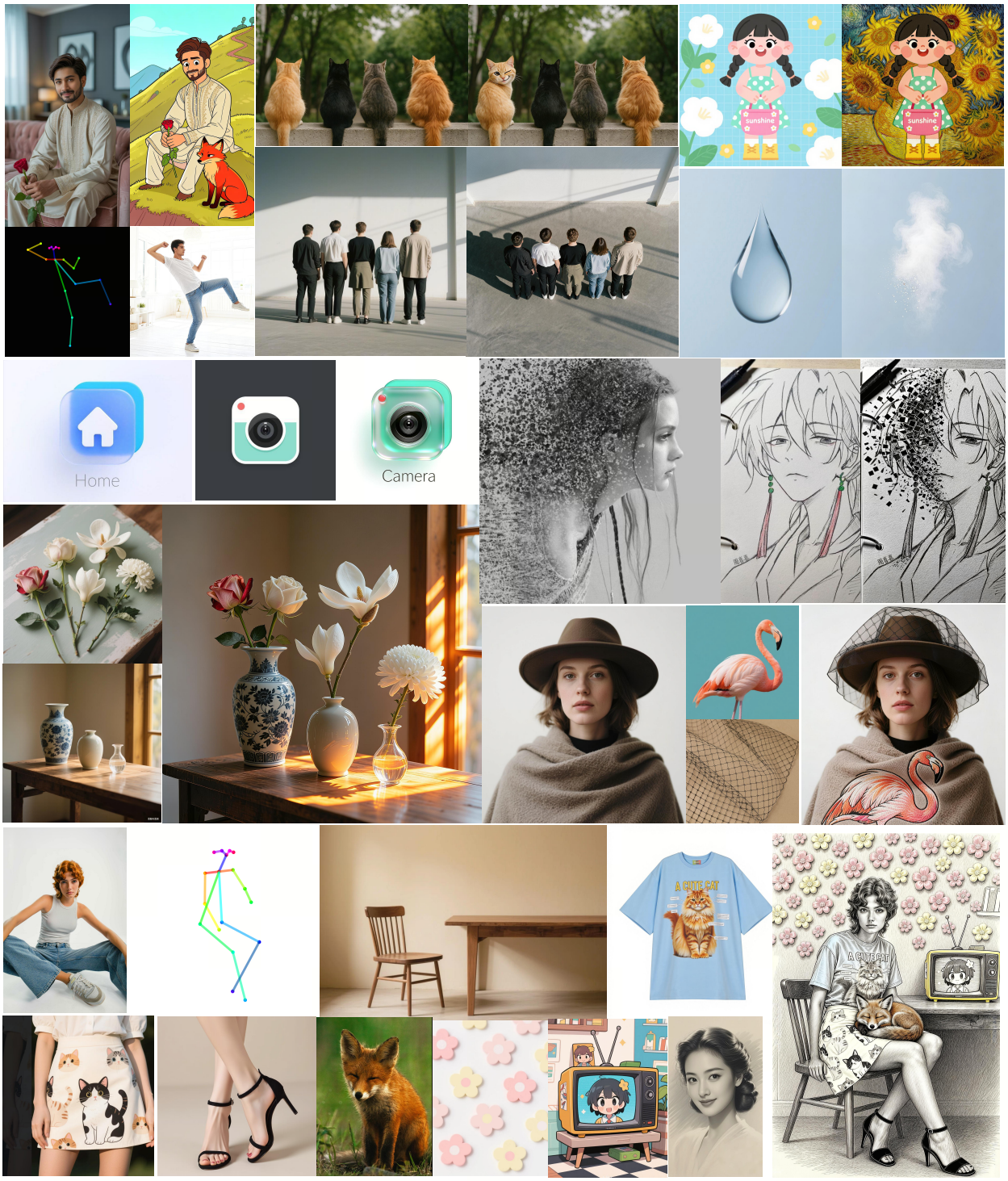


Figure 4 Precise/Reference image editing with MMCORE. Input images are framed at left; output image at right. The model handles multi-image contexts (>10 inputs) with fine-grained localized control.

1 Introduction

Recently, multimodal generation has emerged as a central problem in machine learning. Two model families dominate current practice: autoregressive (AR) models, which underpin state-of-the-art large language models (LLMs) and vision–language models (VLMs), and diffusion or flow-matching (FM) models, which excel at visual generation tasks such as image and video synthesis. To combine semantic reasoning with high-fidelity visual generation, recent unified models integrate FM mechanisms into AR architectures, as exemplified by Transfusion [48] and BAGEL [5].

Directly integrating these paradigms into a single training framework poses significant challenges. In particular, diffusion-based visual modeling degrades training efficiency: unlike text, where a single forward pass supports both understanding and generation, image modeling must alternate between clean features for understanding and noisy features for generation, preventing simultaneous optimization in a single pass [5].

To reduce training costs and better align with LLMs, purely autoregressive visual generation models have been explored [9, 18, 34]. However, their generation quality still lags behind diffusion-based approaches [30], largely due to non-sequential priors and the quantization of visual representations. Other methods, such as MetaQueries [22], decouple the two paradigms by using a frozen multimodal LLM to produce conditioning embeddings and by training a diffusion model via a lightweight connector. This strategy preserves modality-specific strengths and efficiently summarizes long-context information, particularly benefiting bidirectional generative architectures such as MMDiT [7].

In this work, we propose a multi-stage training framework that achieves a favorable trade-off between generation quality and training efficiency. We first fine-tune an MLLM in an autoregressive manner with causal attention to produce compact latent visual embeddings across diverse multimodal data. The latent visual embeddings are trained to align with semantic representations from state-of-the-art (SoTA) vision–language encoders such as SigLIP [35]. In the second stage, we train a diffusion model conditioned on text and the latent visual embeddings using flow matching. By simplifying the adaptation of the MLLM, we eliminate the need for explicit connector modules, resulting in a cleaner architecture that is more amenable to post-training procedures such as supervised fine-tuning (SFT) and reinforcement learning (RL).

We evaluate our approach across a range of scaling settings, including loss configurations, batch sizes, and image resolutions. Empirically, incorporating latent visual embeddings enables the diffusion model to leverage richer visual understanding from the VLM beyond text-only conditioning, leading to consistent improvements over state-of-the-art baselines (Fig. 1, 2). These results demonstrate that high-quality multimodal generation can be achieved by combining autoregressive semantic modeling with diffusion-based visual synthesis, without tightly coupling the two paradigms within a single network.

Finally, our current architecture decouples image understanding and image generation by relying on separate visual encoders. An important next step is to unify these pathways by transforming the latent visual tokens into a sequentialized stream of image-latent tokens suitable for both tasks, enabling both processes to share a single forward pass and further improving model compactness and efficiency.

2 Related Work

2.1 Unified Multimodal Models

Unified Multimodal Models (UMMs) have garnered significant attention for their capacity to support both understanding and generation, demonstrating robust generalization across visual understanding, generation, editing, and various downstream tasks. Early UMMs [3, 33, 39, 43, 45, 48] typically utilized a single Transformer backbone to process interleaved image and text tokens. Subsequent approaches introduced separate branches for understanding and generation to ease optimization and improve performance [5, 31, 42, 44]. However, these models generally rely on large-scale pre-training over massive image–text corpora, creating significant resource barriers that impede broader research accessibility and exploration.

A parallel line of work, including BLIP3-o [1], MetaQueries [22], and UniWorld [19], seeks to construct efficient UMMs by coupling frozen multimodal LLMs with trainable diffusion decoders. While this double-fusion

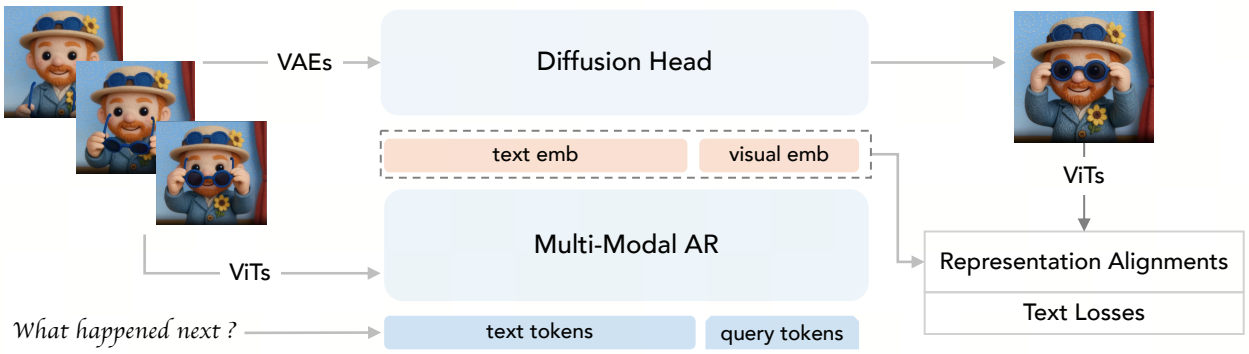


Figure 5 Architecture of MMCORE for low-cost Unified Multi-modal Model. Multimodal information from a VLM is compressed into learned visual latent embeddings, which condition a diffusion-based image generator.

structure shares architectural similarities with the modality-specialized mixture-of-experts designs found in LMFusion [31] and BAGEL [5], the underlying resource requirements differ fundamentally. The latter models often initialize the generation branch with duplicated LLM weights, necessitating orders of magnitude more training tokens and extensive private data. In contrast, LightFusion [40] trains fused base models—derived from publicly available checkpoints—using strictly public data, achieving competitive performance within a highly efficient and accessible training regime.

2.2 Cross-Modality Alignment and Representation Transfer

Cross-modal alignment aims to map heterogeneous visual and textual representations into a shared semantic space that facilitates both understanding and generation. Early dual-encoder frameworks, such as CLIP [25] and ALIGN [13], learned modality-invariant embeddings via contrastive objectives. Later multimodal transformers (e.g., BLIP/BLIP-2 [16, 17], PaLI-X [2]) further advanced this by enabling richer vision–language fusion. To reduce computational costs, recent research has explored representation transfer and lightweight alignment strategies for pretrained models [11, 12, 36]. In generative contexts, diffusion-based models frequently reuse text embeddings to condition image synthesis [24, 26, 28], motivating the direct transfer of multimodal LLM representations to diffusion decoders [1, 19, 22]. Our approach builds upon this foundation, leveraging pretrained multimodal representations and lightweight alignment to enable efficient, high-fidelity image generation without the need for large-scale retraining.

3 Approach

In this section, we introduce our unified framework, detailing the model architecture and training strategies designed to align Multimodal Large Language Models (MLLMs) with diffusion-based generators. Given the significant infrastructure gap between current open-source models and proprietary internal state-of-the-art (SoTA) systems, we conduct all experiments using our internal high-performance training pipeline to ensure rigorous evaluation and make our conclusions transferable.

As illustrated in Fig. 5, our architecture consists of an MLLM at the base, which infers high-level semantic visual latent embeddings from multimodal and query inputs, and a causal diffusion-based network at the top for pixel-level understanding and generation. During training, target images provide dual supervision: serving as pixel-level targets for diffusion denoising and as semantic-level targets for learning visual tokens.

3.1 Preliminaries

MetaQueries [22] addresses the representational mismatch between the autoregressive latent space of MLLMs and the conditioning interfaces required by diffusion decoders [10, 23, 27]. The method introduces a set of learnable query tokens $\mathbf{Q} \in \mathbb{R}^{N \times D}$ appended to the multimodal token stream, where N is the query budget and D is the MLLM hidden dimension. Through self-attention [37], these queries distill task-relevant

multimodal information into a fixed-size representation, which is then projected into the diffusion conditioning space via a lightweight transformer connector. This mechanism enables the low-interference transfer of MLLM knowledge to high-fidelity image synthesis without requiring end-to-end retraining.

However, relying solely on queried embeddings presents two critical limitations:

(i) Inflexible Context Budget. A fixed query budget N is suboptimal for handling variable context complexity: a small N fails to capture the nuances of lengthy prompts, while an excessively large N introduces redundancy and computational overhead.

(ii) Weak Alignment. Supervision derived exclusively from the diffusion objective results in insufficient alignment between the queried representations and the target visual latent space. This leads to reduced training efficiency and heightened sensitivity to data distribution and objective balancing [23, 27].

Empirically, we demonstrate that relying solely on query tokens yields significantly weaker performance compared to utilizing the full context. Furthermore, we observe that employing a frozen MLLM backbone results in substantially inferior generation quality relative to its fine-tuned counterparts.

3.1.1 MLLM for Semantic Visual Prediction

To overcome the limitations of fixed query budgets and weak supervision, we adopt the query-based token generation strategy from MetaQueries but introduce three critical architectural and training modifications. These changes are designed to bridge the modality gap between the MLLM's text-dominant latent space and the spatially aware representations required for high-fidelity generation:

(i) Joint Fine-tuning of the MLLM Backbone. Unlike prior approaches that rely on lightweight adapters while keeping the LLM frozen, we fully fine-tune the multimodal backbone jointly on understanding and generation datasets. This enables the model to adapt its weights specifically for visual token synthesis while retaining the core of its pre-trained knowledge. We acknowledge that this aggressive adaptation currently induces a minor regression in general MLLM understanding capabilities. However, we posit that this is a curriculum scheduling issue rather than an intrinsic architectural limitation; we anticipate that introducing generation objectives earlier in the pre-training stage, combined with model scaling, will effectively resolve this interference in future iterations.

(ii) Semantic Visual Alignment via Distillation. Relying solely on diffusion loss provides a sparse and noisy supervision signal for learning visual tokens. To mitigate this, we introduce explicit intermediate supervision using the latent feature space of a frozen, state-of-the-art vision encoder (e.g., ViT [6] or SigLIP [46]). By regressing the MLLM's generated query tokens toward these semantically dense visual embeddings, we provide a stable, low-variance target. This distillation process acts as a "scaffold," significantly accelerating convergence and ensuring the learned tokens encapsulate robust high-level visual semantics before they are passed to the diffusion head.

(iii) Dual-Pathway Conditioning. A fixed number of query tokens imposes an information bottleneck, potentially discarding dense textual details required for complex instructions. To address this, we implement a dual-pathway conditioning mechanism: we retain the MLLM's original full-sequence text embeddings alongside the learned visual query tokens. This decomposition allows for a specialized division of labor: the visual query tokens capture global semantics and cross-modal grounding, while the variable-length text embeddings preserve fine-grained lexical details and instruction-following constraints.

Formally, let Φ_{vit} be a frozen ViT encoder and $\mathcal{F}_\theta$ be the trainable MLLM. We pool ViT patch features into a $K \times K$ grid to match the query budget ($N = K^2$), yielding target visual features $\mathbf{v} = \Phi_{\text{vit}}(I)$. We optimize the visual tokens via a cosine similarity loss:

$$\mathcal{L}_{\text{vis}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\mathcal{F}_\theta(\mathbf{Q}_i)^\top \mathbf{v}_i}{\|\mathcal{F}_\theta(\mathbf{Q}_i)\| \, \|\mathbf{v}_i\|} \right). \tag{1}$$

The final training objective is:

$$\mathcal{L}_{\text{mllm}} = \lambda_t \mathcal{L}_{\text{llm}} + \lambda_a \mathcal{L}_{\text{vis}}. \tag{2}$$

Empirically, we find that unfreezing the backbone and explicitly aligning visual tokens consistently improves the quality of image generation and instruction-based editing compared to frozen interfaces, without degrading instruction-following capabilities.

3.1.2 MLLM Configurations

Two-stage Training. As indicated in Eq. 2, we do not employ a flow-matching loss for the MLLM, allowing it to be trained independently in a two-stage manner. We observe that $\mathcal{L}_{\text{vis}}$ alone is sufficient to adapt the MLLM for semantic feature prediction. Adding flow matching—which would require computationally expensive joint training—yields only marginal improvements while significantly increasing training costs.

Scaling the Query Budget. We analyze the impact of the visual token count by varying N from 1 to 128. We find that $N = 64$ offers the optimal trade-off between expressivity and efficiency. A smaller N fails to capture sufficient detail for long prompts, while a larger N provides diminishing returns relative to the increased compute. Detailed ablation results are provided in Section 4.

Scaling Batch Size and Sequence Length. To maximize training efficiency, we scale our infrastructure from small to large by increasing the number of GPUs, where the batch token length of the MLLM.is expanded. Following the training of a diffusion head on top of the aligned visual tokens, we observe that performance improves consistently with scale, showing no signs of saturation within our maximum compute budget. This aligns with established scaling laws for large-model training [8, 14].

SFT for Improved Alignment. Our decoupled pipeline facilitates staged optimization. Following large-scale pre-training on noisy image–text data, we apply Supervised Fine-Tuning (SFT) using a curated multimodal instruction dataset. A brief SFT phase (e.g., ∼2K steps) significantly enhances text faithfulness and controllability, consistent with prior findings in instruction tuning [21, 41].

3.2 Modality Mix with Diffusion Head

We leverage a pre-trained Text-to-Image (T2I) Diffusion Transformer (MMDiT) as our generative backbone. Capitalizing on its established text–image alignment capabilities, we extend this model to support multi-reference generation and editing tasks. As demonstrated in prior works [30, 32, 38], pre-trained backbones can be efficiently adapted to new modalities given high-quality fine-tuning data, significantly reducing the computationally prohibitive cost of training from scratch.

Aligning T2I for Interleaved Generation. To handle interleaved multimodal sequences, we re-purpose the model's original self-attention mechanism to support *diffusion-based teacher forcing*, as illustrated in Fig. 6. We implement a block-causal attention mask where the generation of the current image frame is conditioned on:

(1) The VAE latent representations (via MMDiT features) of all preceding images.

(2) The current frame's text and visual latent embeddings.

Crucially, we explicitly **exclude** the semantic visual latent tokens of previous frames from the attention context. Our empirical analysis indicates that attending to historical visual tokens destabilizes the optimization landscape, leading to performance degradation. We attribute this to the complementary nature of the signals: the VAE latents from previous images preserve dense, high-frequency historical details, while the current visual latent tokens provide the necessary high-level semantic guidance for the immediate generation step.

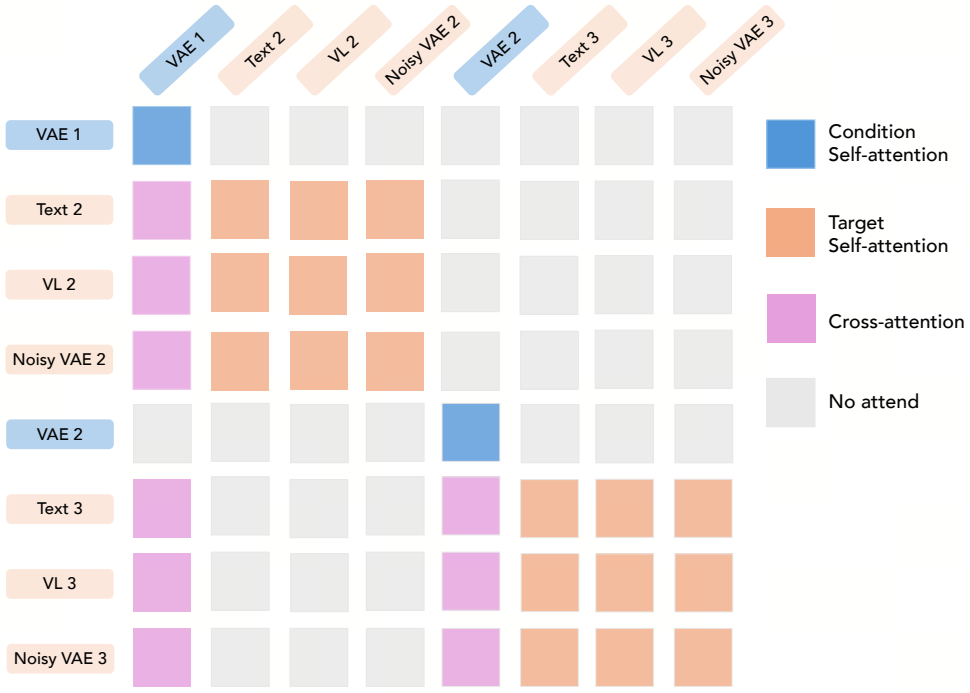


Figure 6 Attention mask for diffusion-head training. Current generation is conditioned on the VAE latents features of all preceding images and the current text/visual latent (**VL**) embeddings.

Efficient Training. Following these architectural modifications, we proceed to a Continued Pre-Training (CPT) stage using a mixture of T2I and interleaved datasets, while keeping the fine-tuned MLLM frozen. To ensure the diffusion model effectively utilizes both conditioning modalities, we employ an **Independent Embedding Dropout** strategy, similar to recent approaches in flow matching [7, 15].

Specifically, we assign independent dropout rates to the text and visual token conditioning pathways. We observe that pre-trained DiT models exhibit a strong inductive bias toward the text conditioning. Thus, we apply a higher dropout rate to text embeddings during the early training stages. This forces the diffusion model to rely more on signals from the MLLM-inferred visual embeddings. This dropout schedule is subsequently annealed to enable robust joint conditioning.

Our streamline integrates multimodal information into the diffusion backbone via meta-query alignment and continued pre-training. In our experiments, we find this strategy is more efficient, achieving performance parity with native unified architectures—such as Transfusion [48] and BAGEL [5]—while requiring only ∼30% of the computational budget typically associated with training these models from scratch.

Finally, keeping the same dropout settings and following the established alignment pipelines [29], we refine the model via Supervised Fine-Tuning (SFT) on a specialized high-quality dataset, followed by Reinforcement Learning with Human Feedback (RLHF). This final stage aligns the model closer to its theoretical upper bound, yielding a network that significantly outperforms the baselines.

Discussion: Necessity of VAEs. Recent advances, such as RAE [47], have demonstrated that discriminative backbones (e.g., ViTs) can potentially support strong image reconstruction. However, current iterations still lag behind VAEs in terms of reconstruction stability and spatial fidelity [1]. Furthermore, the efficacy of these ViT-based approaches has yet to be empirically validated in complex interleaved generation scenarios. We posit that as the reconstruction capabilities of discriminative encoders converge with those of generative autoencoders, the distinction between the two will blur. In such a future regime, the explicit dependency on

[1] https://bfl.ai/research/representation-comparison

VAE latents for historical context—and by extension, the necessity of the diffusion head mechanism—may be eliminated, allowing for a fully unified token-based architecture.

4 Results

In this section, we illustrate the results to demonstrate the effectiveness of MMCORE. Notice for results comparison, we adopted the same prompt input for ours and Seedream 4.0.

4.1 Quantitative Benchmarks

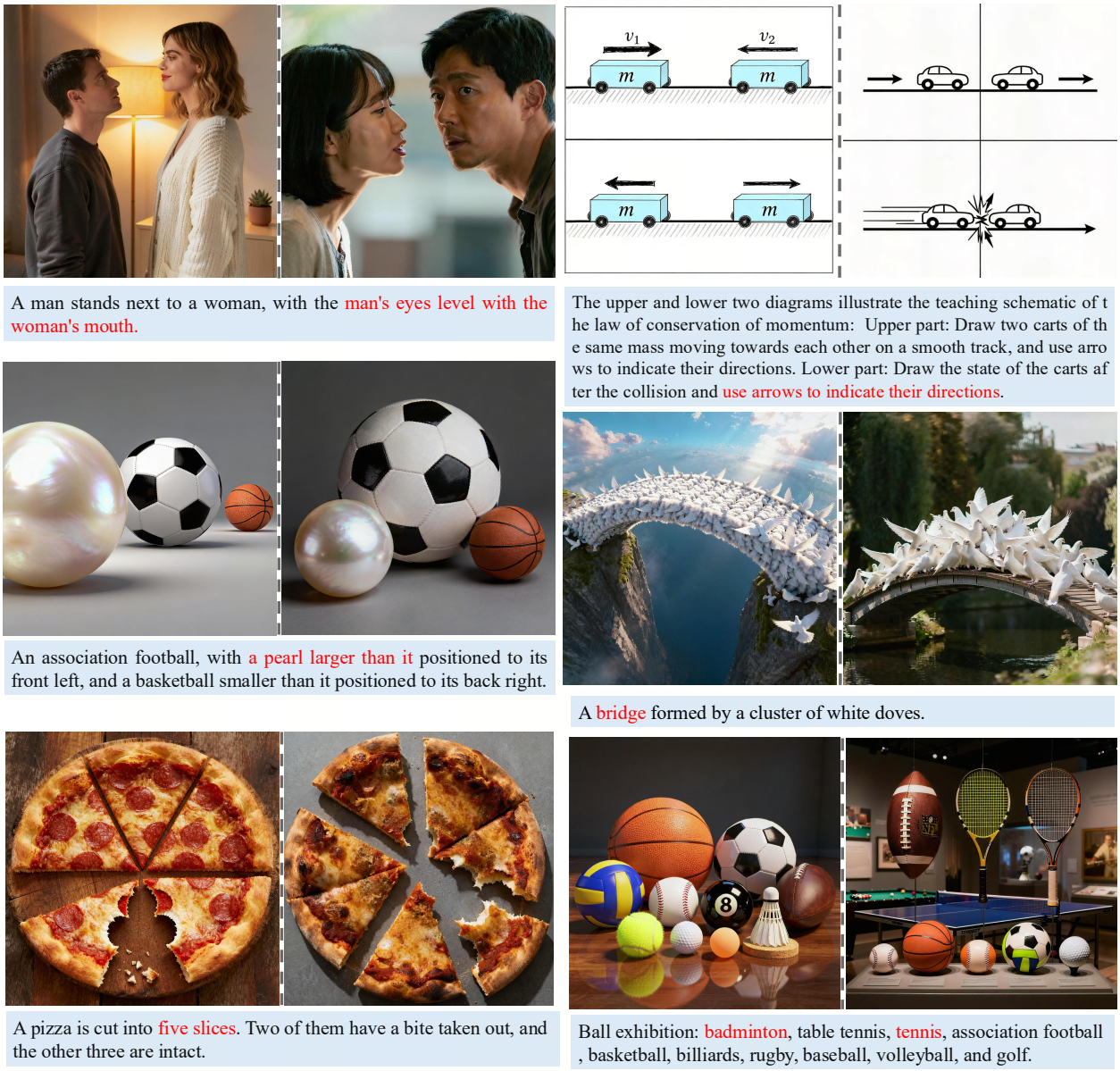


Figure 7 Text-to-image generation comparison against Seedream 4.0. For each pair of images, left: MMCORE; right: Seedream 4.0.

We benchmark MMCORE on an internal suite (DreamBench) [30] designed to evaluate both text-to-image generation and image editing via a robust automated pipeline.

Text-to-Image Generation. As shown in Fig. 1a, MMCORE achieves superior prompt-image coherence based on our automatic evaluation metrics [30], significantly outperforming baseline models including Seedream4.0 [30] and other SoTA methods, thanks to our MLLM's strong text interpolation abilities.

Image Editing. In editing tasks (Fig. 1b,1c), our model also demonstrates enhanced instruction alignment and consistency, validating the effectiveness of the proposed dual-pathway conditioning, where the ViTs indeed provide an additional semantic visual representation compared to purely diffusion based methods with text/meta-info connection as in prior methods [30, 38].

Human Evaluation. To corroborate automated metrics, we conduct a comprehensive human evaluation (Fig. 1) across three axes: (i) Prompt–Image Alignment (English/Chinese), (ii) Structural & Visual Fidelity, and (iii) Editing Consistency (preserving identity and background). MMCORE delivers consistent improvements across all dimensions over our baseline, indicating high reliability in following complex instructions while maintaining visual aesthetics. It also validate the fidelity of our autoeval metrics.

4.2 Qualitative Analysis

Fig. 7 highlights MMCORE's robustness on counterfactual and reasoning-heavy prompts, where baseline models often revert to stereotypical visual priors and lose key textual constraints. Relative to Seedream 4.0, MMCORE demonstrates stronger compositional role binding and attribute grounding: it correctly assigns asymmetric relations (e.g., the man's eye level aligned with the woman's mouth), preserves literal interpretations of non-literal phrases (e.g., "a tree skipping rope"), and faithfully instantiates fine-grained, localized details (e.g., a printed orange cat on a shirt with the tail extending along the sleeve). Beyond single-scene reasoning, MMCORE better adheres to dense, multi-part specifications, such as structured grids/collages, indicating improved text fidelity, spatial reasoning, and constraint satisfaction under complex prompts.

Fig. 4 highlights the model's versatility in precise editing. MMCORE generalizes across a wide spectrum of edit types, ranging from single-image semantic modifications (e.g., viewpoint shifts, style transfer) to complex multi-image compositions. Notably, the model scales effectively to long contexts, maintaining fine-grained control even when conditioned on over 10 input images. Fig. 8 shows MMCORE's improved edit controllability and instruction fidelity over Seedream 4.0. It more consistently performs precise spatial and geometric edits (object relocation, position swaps, shape changes) while better preserving non-target content.

4.3 Ablation Study

To efficiently validate our architectural decisions and training strategies, we employ a lightweight diffusion head for all ablation studies. We evaluate performance using GPT-4o and Doubao-VL[2] as automated judges to assess semantic alignment and visual quality. Specifically, we design a prompt that instructs these models to rate the alignment between the generated image and the text condition, yielding a normalized score between 0 and 1. Here we reveal a few important conclusions from our study.

Architecture and Fine-tuning Strategy. We first established a baseline following the MetaQueries setting [22], coupling a frozen VLM with a pre-trained DiT by replacing the original text encoder. Initially, training the query tokens and connectors using only the DiT loss resulted in poor convergence compared to standard text-to-image diffusion models. To address this, we adopted our visual latent alignment loss (Eq. 2) within a two-stage training framework (Sec. 3.1.2), which significantly accelerated convergence.

With this stable baseline, we examined the capacity required to bridge the modality gap between the MLLM and DiT. As shown in the top section of Table 1, a shallow 2-layer connector yields suboptimal results (0.6791). Increasing the depth to 6 layers provides a substantial performance boost (+10.5%), confirming that a high-capacity projector is essential for aligning disparate latent spaces. We further compared parameter-efficient tuning (LoRA [12]) against full model fine-tuning. While LoRA improves upon the baseline, we found that Full Fine-tuning achieves the lowest convergence loss and superior generation quality (> 0.81), justifying the

[2] https://www.volcengine.com/experience/ark?model=doubao-1-5-thinking-vision-pro-250428

Figure 8 Single- and multi-image editing results comparison against Seedream 4.0.

Table 1 Ablation on MLLM Architecture & Tuning Strategy. Evaluation of connector depth and fine-tuning methods on text-to-image (T2I) generation (Dreambench). Metrics show absolute scores and absolute gains (Δ) compared to the fixed 2-Layer Baseline.

Tuning Strategy	Configuration	GPT-4o ↑	Doubao ↑
1. Connector Depth Analysis (Fixed VLM)			
Baseline (50k)	2 Layers (Ref)	0.6791	0.7276
	3 Layers	0.7490	0.7910
	6 Layers	0.7843	0.8396
	Δ vs. Baseline	**+10.5%**	**+11.2%**
2. Full Model Fine-Tuning			
Full Fine-Tuning (50k)	Full Weights	0.8114	0.8417
	Full Weights (Large BS)	0.8199	0.8528
+ SFT (2k)	**Full Weights (Large BS)**	**0.8585**	**0.8915**
	Δ vs. Baseline	**+25.9%**	**+16.4%**

additional computational cost. Furthermore, scaling the batch size by 5× yielded consistent improvements in text-image alignment, raising the GPT-4o score from 0.8114 to 0.8199.

Effectiveness of Supervised Fine-Tuning (SFT). Finally, we evaluated the efficiency of a targeted SFT phase (Table 1, bottom). We observed that extended pre-training (50k steps) exhibits diminishing returns, with performance plateauing around 0.82. In contrast, transitioning to a short, high-quality SFT phase (2k steps) dramatically boosts the alignment score to **0.8585** (GPT-4o) and **0.8915** (Doubao). This demonstrates that SFT is indispensable for aligning model outputs with granular human esthetic preferences.

Impact of Visual Latents on Conditional VAEs. During diffusion head training, we investigated the potential benefits of augmenting the conditional VAE's DiT encoder features with our learned visual latent embeddings. Leveraging a pre-trained DiT backbone, we found that introducing these additional visual cues resulted in a marked performance degradation. Specifically, our automated evaluation (Doubao for Edit Alignment) revealed a significant drop from 55.2 to 30.62 after 15k fine-tuning steps. Qualitatively, the model exhibited a few failure modes, including severe image artifacts and a tendency to trivially copy reference inputs. These findings indicate that fusing heterogeneous visual features—specifically, dense VAE latents and high-level ViT embeddings—introduces substantial optimization challenges when utilized simultaneously as conditioning signals.

5 Conclusion and Limitations

In this work, we present a resource-efficient, streamlined, yet highly effective framework for aligning the long-context understanding and reasoning capabilities of Multimodal Large Language Models (MLLMs) with latent visual representations. By introducing a trainable query mechanism alongside a dual-pathway conditioning strategy, our approach bridges the modality gap without the prohibitive costs of end-to-end retraining. Extensive experiments demonstrate that our method significantly improves image generation fidelity and instruction adherence compared to baseline counterparts, such as Seedream 4.0 [30], particularly in complex multi-turn interleaved generation scenarios.

Despite these advancements, we admit there is still a performance gap between ours and Nano-Banana-pro [4] / GPT image 1.5 [20]. We think this might be due to the gap in the MLLM model since adding prompt rewrites from SoTA VLM models, MMCORE performs much better. Secondly, our current framework exhibits limitations that highlight critical directions for future research:

The Understanding-Generation Trade-off. A persistent challenge in current Unified Multimodal Models (UMMs) is the performance degradation observed in pure understanding tasks after generative alignment.

Consistent with other SoTA approaches, connecting the MLLM to a generative decoder and fine-tuning for visual synthesis imposes a "tax" on the model's original reasoning capabilities (e.g., VQA, OCR), especially for those highly optimized product models. While our joint training strategy mitigates this catastrophic forgetting, achieving parity with specialized understanding models remains an open challenge.

Redundancy in Visual Latents. Currently, our learned visual tokens function as an *amendment* to text conditioning rather than a comprehensive *supplement* or replacement. This implies a degree of semantic redundancy, where the visual tokens alone are insufficient to drive high-fidelity generation without the auxiliary ViT encoder and diffusion decoder. This limitation underscores the necessity for a next-generation "Omni-Tokenizer"—a unified visual representation capable of supporting both pixel-perfect reconstruction (like a VAE) and high-level semantic reasoning (like a ViT) simultaneously. Developing such a tokenizer to unify understanding and generation at the representation level is the primary focus of our future work.

References

[1] Jiuhai Chen, Zhiyang Xu, Ran Xu, et al. Blip3-o: A family of fully open unified multimodal models–architecture, training and dataset. arXiv preprint arXiv:2505.09568, 2025.

[2] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565, 2023.

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

[4] Google DeepMind. Is nano banana pro a low-level vision all-rounder? a comprehensive evaluation on 14 tasks and 40 datasets. arXiv preprint arXiv:2512.15110, 2025. URL https://arxiv.org/abs/2512.15110.

[5] 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.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In ICLR, 2024.

[8] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. In arXiv preprint arXiv:1706.02677, 2017.

[9] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 15733–15744, 2025.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. NeurIPS, 2020.

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrill, Andrea Corrado, Sergei Vassilvitskii, D Li, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning (ICML), pages 2790–2799. PMLR, 2019.

[12] 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 (ICLR), 2022.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning (ICML), pages 4904–4916. PMLR, 2021.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

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

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (ICML), pages 12888–12900. PMLR, 2022.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning (ICML), pages 19730–19742. PMLR, 2023.

[18] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems, 37:56424–56445, 2024.

[19] Bin Lin, Zongjian Li, Li Yuan, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147, 2025.

[20] OpenAI. Gpt image 1.5 system card. https://platform.openai.com/docs/models/gpt-image-1-5, 2025. Accessed: 2026-01-27.

[21] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155, 2022.

[22] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256, 2025.

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. Proceedings of ICCV, 2023.

[24] 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. arXiv preprint arXiv:2307.01952, 2023.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), pages 8748–8763. PMLR, 2021.

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

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. Proceedings of CVPR, 2022.

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In Advances in Neural Information Processing Systems (NeurIPS), volume 35, pages 36479–36494, 2022.

[29] ByteDance Seed Vision Team. Seedream 2.0: A native chinese-english bilingual image generation foundation model. arXiv preprint arXiv:2503.07703, 2025.

[30] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427, 2025.

[31] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188, 2024.

[32] Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. arXiv preprint arXiv:2411.06686, 2024.

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

[34] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. arXiv preprint arXiv:2404.02905, 2024.

[35] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025. doi: 10.48550/arXiv.2502.14786.

[36] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In Advances in Neural Information Processing Systems (NeurIPS), volume 34, pages 200–212, 2021.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In NeurIPS, 2017.

[38] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. arXiv preprint arXiv:2506.05083, 2025.

[39] 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.

[40] Zeyu Wang, Zilong Chen, Cihang Xie, et al. Lightfusion: A light-weighted, double fusion framework for unified multimodal understanding and generation. arXiv preprint arXiv:2510.22946, 2025.

[41] Jason Wei, Maarten Bosma, Vincent Zhao, et al. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2022.

[42] Yichen Wei, Wei Shen, Yang Liu, Yahui Zhou, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. arXiv preprint arXiv:2504.16656, 2025.

[43] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yuxuan Ma, Xingchao Liu, Zizheng Pan, Wenbo Chang, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

[44] Shitao Wu, Kai Zheng, Fenging Zhang, Yimin Wang, Han Zhang, Yifan Zhang, Yu Zhou, Wei Feng, Yan Liu, et al. Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871, 2025.

[45] 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. In ICLR, 2025.

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In International Conference on Computer Vision (ICCV), pages 11975–11986, 2023.

[47] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690, 2025.

[48] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039, 2024.

Appendix

A Ethical Claims

The images presented in the paper are from our lisenced ones, and public license-free websites such as Unsplash and Pixabay. In addition, note that the technique proposed in this paper aims to facilitate the user's common tasks that are widely demanded in industry for ethical purposes. It SHOULD NOT be applied to unwanted scenarios such as generating violent and sexual content. It might also inherit the biases and limitations of T2I models. Therefore, we believe that the images or models synthesized using our approach should be carefully examined and presented as synthetic.