Symbiotic-MoE: Unlocking the Synergy between Generation and Understanding

Xiangyue Liu[1], Zijian Zhang[2], Miles Yang[2], Zhao Zhong[2],
Liefeng Bo[2], Ping Tan[1⋆]

[1]HKUST [2]Tencent Hunyuan

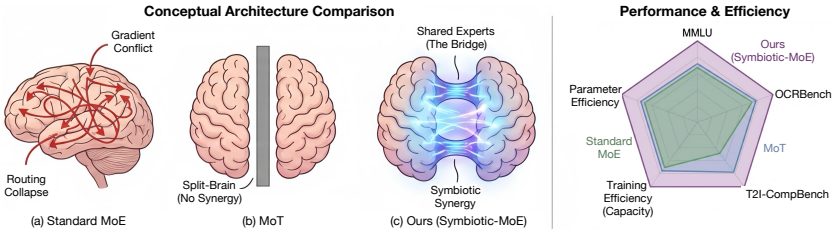


Fig. 1: Teaser. We visualize the evolution of training paradigms: moving from (a) destructive interference and (b) the "Split-Brain" compromise to (c) symbiotic synergy. As shown in the radar chart, our framework achieves holistic superiority, simultaneously boosting generation and understanding capabilities with zero-parameter overhead and maximal parameter efficiency.

Abstract. Empowering Large Multimodal Models (LMMs) with image generation often leads to catastrophic forgetting in understanding tasks due to severe gradient conflicts. While existing paradigms like Mixture-of-Transformers (MoT) mitigate this conflict through structural isolation, they fundamentally sever cross-modal synergy and suffer from capacity fragmentation. In this work, we present Symbiotic-MoE, a unified pre-training framework that resolves task interference within a native multimodal Mixture-of-Experts (MoE) Transformers architecture with zero-parameter overhead. We first identify that standard MoE tuning leads to routing collapse, where generative gradients dominate expert utilization. To address this, we introduce Modality-Aware Expert Disentanglement, which partitions experts into task-specific groups while utilizing shared experts as a multimodal semantic bridge. Crucially, this design allows shared experts to absorb fine-grained visual semantics from generative tasks to enrich textual representations. To optimize this, we propose a Progressive Training Strategy featuring differential learning rates and early-stage gradient shielding. This mechanism not only shields pre-trained knowledge from early volatility but eventually transforms generative signals into constructive feedback for understanding. Extensive experiments demonstrate that Symbiotic-MoE achieves rapid generative convergence while unlocking cross-modal synergy, boosting inherent understanding with remarkable gains on MMLU and OCRBench.

Keywords: Mixture-of-Experts Transformers · Unified Generation and Understanding · Catastrophic Forgetting

⋆ Corresponding author.

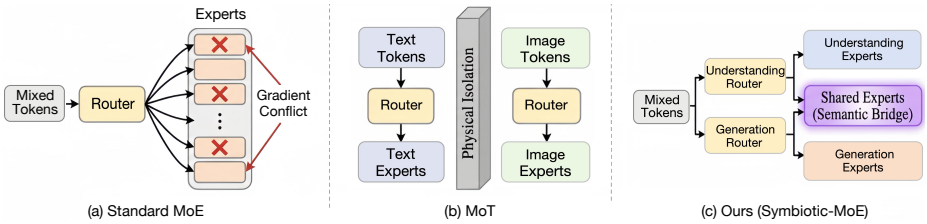


Fig. 2: Comparison of Architectures. (a) Standard MoE suffers from routing collapse due to multi-task gradient conflicts (like cognitive dissonance). (b) MoT avoids conflict via physical isolation but induces a "Split-Brain" dilemma, which structurally hinders cross-modal synergy and knowledge transfer. (c) Our Symbiotic-MoE introduces shared experts as a semantic bridge, enabling co-evolution of generation and understanding within a unified, parameter-efficient architecture.

1 Introduction

The convergence of perception and creation within a single cognitive system has long been a holy grail in the pursuit of Artificial General Intelligence (AGI). In recent years, Large Multimodal Models (LMMs) have made remarkable strides in understanding the visual world [1, 25, 38]. However, enabling these models to also generate visual content that transforms them into true "Any-to-Any" omni-modal foundation models remains a formidable challenge. Recent advances have attempted to unify vision and language through diverse paradigms, ranging from discrete tokenization [35, 37] to continuous feature alignment [3, 48, 57]. Despite these architectural innovations, a critical optimization bottleneck persists: extending a pre-trained understanding model with generative capabilities often triggers severe catastrophic forgetting.

This phenomenon stems from the intrinsic stability-plasticity dilemma [30, 32]. Visual understanding tasks, which require the model to converge onto precise semantic representations (many-to-one mapping), fundamentally conflict with the high-entropy, divergent nature of generative tasks (one-to-many mapping). When naively co-trained, the high-variance gradients from the generative objective tend to overwhelm the established optimization landscape of the understanding task, washing away pre-trained knowledge. Consequently, current unified models often face a zero-sum game: significant improvements in generation quality typically come at the expense of understanding retention, or necessitate complex, multi-stage training strategies to mitigate interference.

To mitigate this interference, prior arts have predominantly resorted to structural isolation. Approaches like Mixture-of-Transformers (MoT) [7,21] or adapter-based methods [8, 45, 49] physically decouple the parameters for different modalities. While effective at preserving stability, this "divide-and-conquer" strategy comes at a steep cost: it often introduces significant parameter overhead, increases inference latency, and most critically, severs the semantic connectivity between modalities. By segregating the processing pathways, these methods forfeit the potential synergy where generative feedback could refine understanding—a cognitive mechanism inherent in biological intelligence.

In this work, we ask a fundamental question: *Can we achieve the stability of isolated architectures while retaining the synergistic benefits of a unified model, without any parameter overhead?* To answer this, we present Symbiotic-MoE, a novel pre-training framework that harmonizes generation and understanding within a native sparse Mixture-of-Experts (MoE) Transformers architecture.

Our key insight is that task interference is not a failure of the unified architecture itself, but a failure of routing dynamics. We observe that standard MoE training leads to routing collapse, where generative tokens monopolize the experts, starving the understanding task. To resolve this, we propose Modality-Aware Expert Disentanglement. Instead of physical separation, we logically partition the expert space into task-specific groups based on their pre-trained activation patterns. Crucially, we introduce shared experts to act as a multimodal semantic bridge. This private-shared design allows task-specific experts to specialize in conflicting objectives, while shared experts facilitate deep cross-modal alignment, turning the conflict into symbiosis.

Furthermore, we recognize that architecture alone is insufficient to tame the volatile training dynamics of generative adaptation. We thus introduce the Progressive Training Strategy. We devise a Knowledge-Inherited Initialization to eliminate the cold-start penalty and a Differential Learning Rate schedule to balance the update pace of different modules. Most notably, we propose a Warmup Gradient Shielding mechanism. This strategy temporarily blocks noisy generative gradients from updating shared experts during the early unstable phase, protecting the foundational knowledge until the generative module matures.

Our contributions are summarized as follows:

- We propose Symbiotic-MoE, a zero-overhead framework that resolves routing collapse in unified multimodal pre-training. It introduces Modality-Aware Expert Disentanglement to resolve resource contention and incorporates Shared Experts as a multimodal bridge to enable seamless cross-modal alignment.
- We introduce a Progressive Training Strategy that harmonizes the conflicting optimization dynamics. By employing Differential Learning Rates and Warmup Gradient Shielding, we effectively navigate the stability-plasticity dilemma, protecting pre-trained knowledge while unlocking generative capabilities.
- Extensive experiments demonstrate that our method not only achieves rapid generative convergence but also boosts understanding capabilities. This provides empirical validation for the symbiotic hypothesis: generative training, when properly orchestrated, can reciprocally enhance multimodal understanding.

2 Related Work

2.1 Unified Multimodal Understanding and Generation

The evolution of Large Multimodal Models (LMMs) has progressed from perception centric systems [5,27,38,41] to unified "Any-to-Any" frameworks [36,46].

Early attempts cascaded LLMs with diffusion models [19, 44], limiting end-to-end synergy. To achieve native unification, pioneers like Chameleon [37] and Emu3 [43] adopted discrete tokenization, formulating image generation as autoregressive next-token prediction. Conversely, continuous approaches like Transfusion [57] and Show-o [49] integrated diffusion or flow-matching objectives directly into the transformers backbone to improve fidelity. More recently, to mitigate modality interference, Janus [45] explored decoupled visual encodings, while OmniGen [48] pushed for a unified diffusion transformer. Despite these strides, training a single backbone for both tasks remains unstable. The divergent nature of generation (one-to-many) and the convergent nature of understanding (many-to-one) create a fundamental optimization paradox [30, 32]. This conflict often leads to a "seesaw effect", where improving generative plasticity inevitably degrades understanding stability.

2.2 Multimodal Mixture-of-Experts (MoE)

Sparse Mixture-of-Experts (MoE) [10, 20, 34] Transformers has emerged as a promising solution to scale model capacity without inflating inference costs, widely adopted in LLMs [9, 12, 29, 58], most notably exemplified by milestones such as Mixtral [18], DeepSeek-MoE [6] and Qwen3 [51]. In the multimodal domain, sparse architectures have proliferated to handle high-resolution inputs efficiently [28, 39, 40, 50]. Beyond foundational works like MoE-LLaVA [22], V-MoE [33], and MM1 [31], recent advances including DeepSeek-V3 [24], MM1.5 [53] further demonstrate the scalability of experts in processing complex modalities. However, the application of MoE in unified generation and understanding remains underexplored. Standard routing mechanisms often fail in this context because the gradients from generative losses are significantly larger and noisier than those from understanding losses. This leads to a routing collapse phenomenon, where experts are overwhelmingly co-opted by the generative task, leaving the understanding capability starved of capacity. Our work is among the first to address this imbalance within a native MoE framework, turning the sparsity from a scaling tool into a mechanism for task disentanglement.

2.3 Task Interference and Resolution Strategies

Catastrophic forgetting [2, 4, 54, 56] arising from gradient conflicts remains a primary bottleneck in extending VLMs with generative capabilities. Prior works mitigate this via two dominant paradigms: Additive Adaptation and Structural Isolation. Additive approaches freeze the pre-trained backbone and append auxiliary modules, such as complex adapters or side-networks [11, 16, 46, 52, 55]. While effective for stability, they intrinsically limit deep cross-modal interaction and incur non-negligible inference latency. Conversely, Structural Isolation methods, exemplified by MoT and recent sparse variants [7, 21, 42], physically decouple processing pathways for different modalities. Although this avoids interference, it enforces a "Split-Brain" architecture that severs potential synergy and suffers

from capacity fragmentation. In contrast, Symbiotic-MoE proposes a Modality-Aware Expert Disentanglement strategy. We achieve the best of both worlds: the specialized optimization of decoupled architectures without parameter overhead, and the deep fusion benefits of a unified model via shared experts. By orchestrating gradient flow rather than physically separating parameters, we resolve interference while fostering positive cross-modal transfer.

3 Method

In this section, we present **Symbiotic-MoE**, a unified pre-training framework designed to harmonize the conflicting objectives of visual generation and multimodal understanding within a single sparse architecture. Unlike prior approaches such as MoT [7, 21] that enforce strict structural isolation, our method operates on the principle of *co-evolution*.

As illustrated in Fig. 3, we address the catastrophic forgetting and routing collapse issues through two core components: (1) **Symbiotic-MoE Architecture** (Sec. 3.2), which employs *Modality-Aware Expert Disentanglement* to partition experts into task-specific groups, bridged by *Shared Experts* for cross-modal alignment, alongside *Knowledge-Inherited Initialization* for a zero-cold-start transition; and (2) **Progressive Training Strategy** (Sec. 3.3), which resolves the stability-plasticity dilemma by orchestrating update dynamics through *Differential Learning Rates* and *Warmup Gradient Shielding*.

3.1 Preliminaries

Sparse Mixture-of-Experts (MoE). We adopt the standard sparse MoE formulation where the dense Feed-Forward Network (FFN) in a Transformer block is replaced by a set of experts $\mathcal{E} = \{E_i\}_{i=1}^{N}$. For an input $\mathbf{x}$, the router selects top-k experts (denoted by indices $\mathcal{I}$) based on gating scores. The output is the weighted sum of the selected experts:

$$\mathbf{y} = \sum_{i \in \mathcal{I}} g_i(\mathbf{x}) \cdot E_i(\mathbf{x}) \tag{1}$$

where $g_i(\mathbf{x})$ is the softmax-normalized routing weight for the chosen expert i. This architecture allows the model to scale capacity while maintaining constant inference costs.

Conflicts in Co-training. Naively integrating generative objectives into a pre-trained MoE-based VLM precipitates a severe stability-plasticity dilemma [30, 32]. The aggressive, high-magnitude gradients required for learning pixel-level synthesis inevitably overwhelm the converged optimization landscape of the understanding task. This gradient interference causes experts to drift uncontrollably toward the generative manifold, erasing pre-trained semantic structures and triggering catastrophic forgetting. Detailed visualization of the conflicts in Supplementary Material Sec. A.

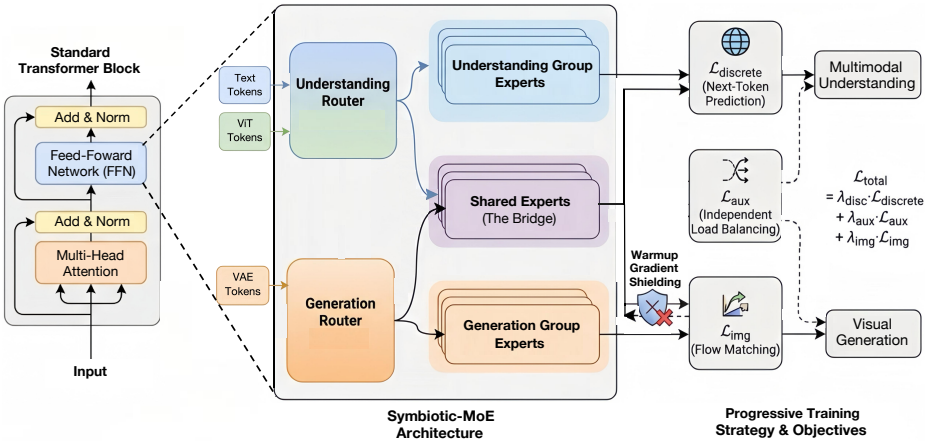


Fig. 3: Overview of Symbiotic-MoE. We re-architect the Transformer FFN layer to resolve task interference. Our framework features: (1) *Modality-Aware Expert Disentanglement*, where input tokens are routed to specialized Understanding (blue) or Generation (orange) experts via decoupled routers; (2) A *Shared Expert Bridge* (purple), which processes all tokens to enforce semantic alignment and prevent modal isolation. The entire framework is optimized via *Progressive Training Strategy*, harmonizing generation and understanding within a unified Transformer backbone.

3.2 Symbiotic-MoE Architecture

Modality-Aware Expert Disentanglement. To resolve the conflict between retaining pre-trained knowledge and learning new generative capabilities, we propose a specialized grouping strategy derived from the intrinsic activation patterns of the experts.

Expert Role Analysis. Instead of random grouping, we conduct a data-driven analysis to identify the role of each expert in the pre-trained VLM. Specifically, we perform inference on two representative benchmarks: MMLU [13] (for linguistic Text tokens) and OCRBench [26] (for discriminative ViT tokens). We calculate the activation frequency of all 128 experts across all 47 MoE layers. By ranking experts based on their cumulative activation rates for Text and ViT tokens, we identify the core experts that are essential for the model's fundamental capabilities by modality. Detailed visualizations of activation landscapes are provided in Supplementary Material Sec. A.

Hard Routing Strategy. Based on the frequency profiling, we implement a hard routing split. For each layer, we designate the top 96 experts with the highest relevance to original modalities (Text and ViT) as the *understanding group*, ensuring that the vast majority of pre-trained knowledge is preserved (stability). The remaining 32 experts in each layer, which are less frequently activated, are repurposed as the *generation group* assigned to the VAE features (plasticity). During training, tokens are routed to their respective groups based on their modality type. Accordingly, the parameters of the original router are sliced and assigned to initialize the specific routers for each group, ensuring a warm start.

The Bridge: Shared Experts. A critical distinction of our Symbiotic-MoE from prior hard-routing methods (*e.g.*, MoT [7, 21]) is the preservation of shared experts. We retain the shared experts in each layer to process *all* tokens, regardless of whether they are Text, ViT, or VAE. This design allows the shared experts to function as a multimodal bridge, facilitating information exchange between the decoupled expert groups. It ensures that the generative process remains semantically aligned with the understanding representations, enabling cross-modal synergy rather than isolation.

Knowledge-Inherited Initialization. Transitioning from a unified MoE to a disentangled architecture typically incurs a cold-start penalty if the new components are randomly initialized. To mitigate this and ensure a seamless adaptation, we propose a *Knowledge-Inherited Initialization* strategy that transfers the learned routing priors and expert capabilities from the pre-trained VLM to our Symbiotic-MoE.

Expert and Shared Weight Inheritance. Since our architecture retains the physical structure of the experts, we directly initialize the parameters of the partitioned understanding and generation groups using the weights from their corresponding expert indices in the original VLM. Similarly, the shared experts, which serve as the multimodal bridge, inherit their parameters directly from the pre-trained shared experts. This ensures that the foundational knowledge stored within the VLM is preserved intact.

Router Weight Slicing. The most critical challenge lies in initializing the two newly decoupled routers (one for understanding group, one for generation group) from the single original router. Random initialization would destroy the learned mapping between tokens and their preferred experts, leading to immediate performance collapse. To address this, we introduce *Router Weight Slicing*. Let $\mathbf{W}_r \in \mathbb{R}^{d \times N}$ denote the projection matrix of the original router, where $N = 128$ is the total number of experts. For a specific group $g \in \{\text{und}, \text{gen}\}$ containing a subset of expert indices $\mathcal{I}_g$, we construct the new router weight $\mathbf{W}_r^g$ by slicing the columns of $\mathbf{W}_r$ corresponding to $\mathcal{I}_g$. Mathematically, $\mathbf{W}_r^g = \mathbf{W}_r[:, \mathcal{I}_g]$. This operation effectively preserves the routing priors: a token that originally preferred Expert i will still produce a high gating score for Expert i in the new sub-router. This strategy achieves a zero-cold-start, allowing the model to maintain near-original understanding performance at iteration zero, as in Table 2.

3.3 Progressive Training Strategy

Even with a disentangled architecture, the simultaneous optimization of a converged VLM and a newly added generative module presents a significant challenge due to their disparate optimization landscapes. To harmonize these conflicting dynamics, we propose a progressive training strategy.

Differential Learning Rates. A unified learning rate is suboptimal for the Symbiotic-MoE due to the maturity gap between modules. The pre-trained VLM components (Text/ViT experts and routers) reside in a sharp local minimum

where a high learning rate (*e.g.*, $1e^{-4}$) triggers immediate divergence and catastrophic forgetting. Conversely, the newly initialized generative components (VAE experts and routers) require a larger step size to escape their initial random state and learn effective representations. To resolve this conflict, we implement a Differential Learning Rate schedule. We assign a larger learning rate ($1e^{-4}$) exclusively to the generation-centric components (VAE experts and routers) to accelerate their convergence. Simultaneously, we apply a conservative learning rate ($1e^{-6}$) to the understanding-centric components (Text/ViT experts, routers, and shared experts) to perform fine-grained adaptation without destroying their pre-trained priors. This differential optimization ensures that each module evolves at its appropriate pace.

Warmup Gradient Shielding. Initial phase of co-training is highly volatile. The generative module produces high variance gradients that can catastrophically distort the well-tuned semantic features within the shared experts. To prevent this, we introduce a temporal gradient shielding mechanism. Specifically, during the warmup period, while VAE tokens are permitted to forward-propagate through shared experts to leverage pre-trained features, we enforce a stop-gradient operation on the backward pass. This strategy acts as a protective buffer, ensuring that shared experts are updated solely by stable Text and ViT gradients while the generative module is in its chaotic early learning phase. Crucially, this shielding is transient. Once the generative optimization trajectory stabilizes after warmup iterations, we remove the constraint to enable full bidirectional gradient flow, thereby facilitating deep cross-modal fusion through shared experts. Concurrently, to counterbalance the update magnitude disparity arising from our differential learning rates ($1e^{-4}$ *vs.* $1e^{-6}$), we apply a gradient scaling factor of 0.1 to generative tokens within the shared experts, preventing aggressive generative signals from monopolizing the shared experts semantic space.

Objective Functions. Our training objective is designed to strictly preserve the original VLM optimization landscape while integrating generative capabilities. The final loss function extends the original VLM objectives by incorporating the image generation term:

$$\mathcal{L}_{total} = \lambda_{disc} \cdot \mathcal{L}_{discrete} + \lambda_{aux} \cdot \mathcal{L}_{aux} + \lambda_{img} \cdot \mathcal{L}_{img} \tag{2}$$

where we set $\lambda_{disc} = 1.0$, $\lambda_{aux} = 0.01$, and $\lambda_{img} = 1.0$ in all of our experiments. The components are defined as follows:

- $\mathcal{L}_{discrete}$ (Inherited): The standard cross-entropy loss for next-token prediction. Specifically, this unifies pure language modeling, multimodal understanding, and conditional text prompts for image generation, ensuring model maintains robust instruction following capabilities.
- $\mathcal{L}_{aux}$ (Inherited): The auxiliary load-balancing loss is to encourage uniform token distribution, which prevents expert collapse and ensures efficient routing. Crucially, consistent with our architectural disentanglement, this loss is computed independently for the understanding and generation groups. This

ensures balanced expert utilization within each modality-specific subspace, rather than enforcing a global balance that might dilute modality specialization.
– $\mathcal{L}_{img}$ (New): The flow matching loss on VAE latents, responsible for aligning the visual features with the generative manifold to synthesize high-fidelity images.

4 Experiments

4.1 Experimental Setup

Experimental Setup. All experiments are conducted on a large-scale proprietary corpus spanning T2I, T2I-Long, LM, and MMU tasks. We adopt a holistic evaluation protocol to rigorously assess both modalities.

Evaluation Metrics. Due to space constraints, the main paper focuses on two representative benchmarks for understanding: **MMLU** [13] (general knowledge reasoning) and **OCRBench** [26] (fine-grained visual perception). For generation, we report FID [15], CLIPScore [14], and HPSv2 [47] on **COCO-30K** [23] to evaluate generation fidelity, alongside **T2I-CompBench** [17] for semantic alignment. Extensive evaluations on a broader suite of multimodal understanding and generation benchmarks are provided in Supplementary Material Sec. B. These extensive results further corroborate the consistent superiority of our method across diverse domains.

Baselines. We compare against two representative architectures under identical data settings: (1) **Standard MoE**: naive unified co-tuning; and (2) **MoT**: structural expert partitioning (96/32 split) *without* shared experts. Crucially, all baselines undergo full fine-tuning to isolate the impact of our architectural and training innovations. *Note: Since the generative module is trained from scratch in this pre-training phase, our primary focus is on architectural comparison and training dynamics rather than absolute aesthetic perfection.*

Implementation Details. All models are initialized from the state-of-the-art Hunyuan-A3B (total 30B parameters) VLM. Training is conducted on 256 NVIDIA H20 GPUs with a global batch size of 2500 sample (∼2M tokens) per iteration, optimized via AdamW with 500 warmup steps. The data mixture is fixed at *T2I:T2I-Long:LM:MMU* = 3 : 3 : 2 : 2. We enforce strict determinism by fixing all random seeds to ensure reproducibility and verified that the setup yields identical training loss curves and evaluation metrics across multiple independent runs. Comprehensive dataset details and hyperparameters are provided in Supplementary Material Sec. C.

4.2 Main Results

Image Generation. We evaluate generative fidelity and semantic alignment on *COCO-30K* (FID, CLIP, HPSv2) and *T2I-CompBench (color, shape, texture)*.

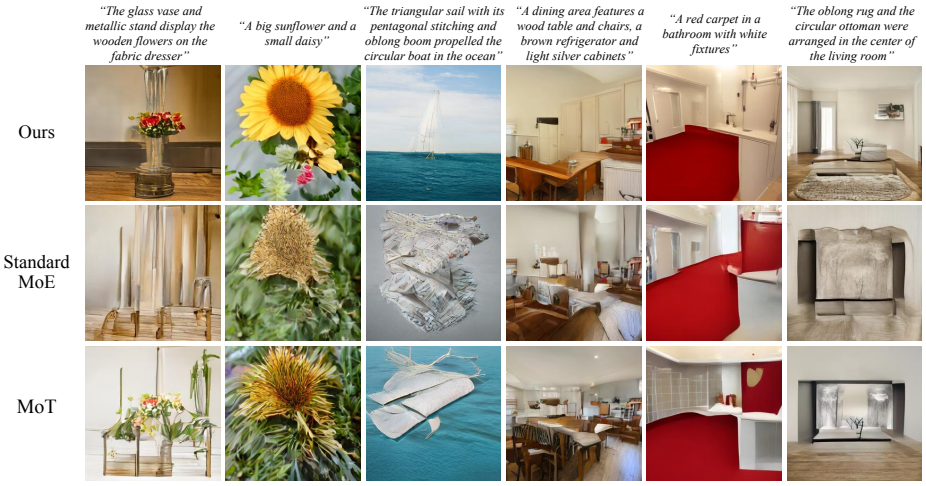
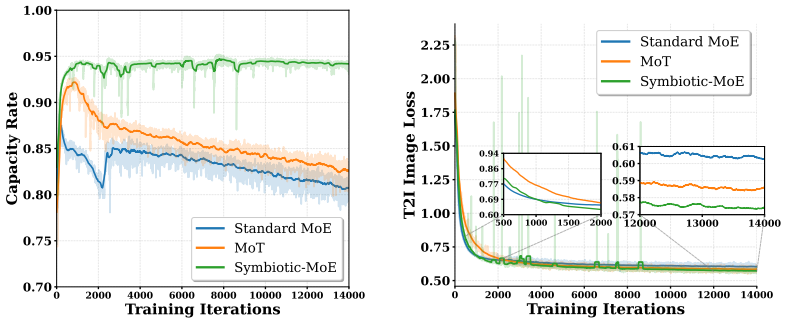


Fig. 4: Visualization Comparisons. We compare samples generated by Symbiotic-MoE (top), Standard MoE (middle), and MoT (bottom). Standard MoE suffers from severe structural collapse and visual artifacts due to gradient conflicts. While MoT recovers basic object shapes, it often misses fine-grained semantic details. In contrast, our method achieves superior fidelity and precise semantic alignment (*e.g.*, *triangular sail*, *oblong rug*), validating the effectiveness of our symbiotic strategy.

Table 1: Quantitative Results. We evaluate **Image Generation** (left) and **Understanding Capabilities** (right). Standard MoE suffers from severe catastrophic forgetting in understanding tasks. MoT and Bagel (the same as MoT except for using independent QKVs) preserve original capabilities via isolation but lack synergy. Symbiotic-MoE (Ours) delivers superior generation quality while synergistically boosting understanding capabilities (*e.g.*, MMLU, OCRBench) beyond the Only_LM_MMU baseline. Best results are highlighted in **bold**. Additionally, we provide results of Symbiotic-MoE trained for 100k steps (in gray) to showcase its extended scaling capacity.

Method	Training Steps	Image Generation Capabilities							Understanding	
		T2I-Comp↑	*Color*↑	*Shape*↑	*Texture*↑	FID↓	CLIP↑	HPSv2↑	MMLU↑	OCRBench↑
Only_LM_MMU	100k	-	-	-	-	-	-	-	0.405	662
Standard MoE	100k	0.36	0.38	0.21	0.40	26.27	0.27	0.19	0.308	571
MoT	100k	0.43	0.51	0.27	0.48	19.87	0.28	0.21	0.392	583
Bagel	100k	0.45	0.52	0.29	0.51	18.42	0.29	0.21	0.396	590
Symbiotic-MoE	100k	**0.49**	**0.55**	**0.35**	**0.56**	**13.65**	**0.31**	**0.23**	**0.507**	**768**

As detailed in Table 1, Standard MoE suffers from routing collapse, resulting in severe artifacts and a high FID of 26.37. While MoT and Bagel [7] improve stability via isolation, it lacks the semantic guidance required for complex prompts, resulting 0.43 and 0.45 respectively on T2I-Compbench [17]. In contrast, our Symbiotic-MoE achieves best performance across all metrics. We ascribe this superior fidelity to the *holistic synergy* of our architecture: the specialized *Generation Group* effectively captures high-frequency visual details, while the *Shared Expert bridge* anchors these details to robust semantic representations. Qualitative comparisons in Fig. 4 corroborate this. As highlighted in the 3rd column, our



(a) Expert Capacity Rate **(b)** T2I Training Loss

Fig. 5: Analysis of Training Dynamics. (a) Standard MoE (blue) and MoT (orange) suffer from routing collapse, indicated by the dropping capacity rate (ratio of non-dropped tokens). In contrast, our Symbiotic-MoE (green) maintains highest expert utilization (∼0.95). **(b)** This structural stability translates into superior optimization efficiency, with our method achieving lower convergence generation loss than baselines.

model accurately synthesizes geometrically complex objects like the "*triangular sail*", whereas baselines struggle with shape consistency. Similarly, in the 6$^{\text{th}}$ column, Symbiotic-MoE successfully disentangles spatial relationships (placing the "*oblong rug*" correctly), avoiding the structural distortion observed in MoT. Additional qualitative results are provided in Supplementary Material Sec. D.

Understanding Capabilities. To verify the understanding ability, we report results on MMLU [13] (reasoning) and OCRBench [26] (perception) in Table 1. Standard MoE succumbs to severe gradient conflict, suffering catastrophic forgetting with MMLU scores plummeting to 0.308. While MoT and Bagel mitigate this decay via physical isolation, they merely maintain a baseline level of 0.392 and 0.396 respectively, failing to leverage the visual signals from the generative stream. Crucially, Symbiotic-MoE achieves a breakthrough. It not only outperforms both standard MoE, MoT, and Bagel baselines by a significant margin but, most notably, surpasses *Only_LM_MMU*—which was trained identically to Symbiotic-MoE but without T2I to isolate the data domain shifts that inherently degrade the VLM. Specifically, our method boosts MMLU performance from 0.405 to 0.507 (+20.1% relative gain) and OCRBench from 662 to 768 (+13.8%). This empirical evidence challenges the prevailing view that generation and understanding are zero-sum competitors. Instead, it validates our core hypothesis: pixel-level generative training, when properly orchestrated via our shared-expert bridge, acts as a powerful fine-grained visual regularizer. It forces the visual encoder to capture denser semantic details required for reconstruction, which reciprocally enhances the model's discriminative reasoning capabilities.

Training Efficiency. Beyond final performance, we analyze the training dynamics to uncover the source of our model's efficiency. As illustrated in Fig. 5a, standard MoE suffers from severe routing collapse, where expert utilization (capacity rate) drops significantly. In contrast, Symbiotic-MoE maintains an exceptionally high capacity rate, approaching **0.95**, a value that significantly exceeds

Table 2: Ablation on Architectural Designs. We evaluate knowledge retention at iteration zero to validate structural integrity. "Tripartite" denotes separating Text, ViT, and VAE; "Bimodal" merges Text/ViT into a single group. The 96/32 Bimodal split with shared experts offers the optimal trade-off.

Grouping Strategy	Expert Allocation (Count)			Shared	Zero-Shot Metrics	
	Text	ViT	VAE	Expert	MMLU	OCRBench
Original VLM	128 (Entangled)		-	✓	0.697	845
– w/o Shared Expert	128 (Entangled)		-	×	0.460	498
Tripartite Split	32	32	64	✓	0.244	109
	44	42	42	✓	0.234	179
Bimodal Split	32 (Understanding)		96	✓	0.285	532
	64 (Understanding)		64	✓	0.453	742
	86 (Understanding)		42	✓	0.561	786
	96 (Understanding)		**32**	✓	**0.601**	**807**

the typical equilibrium threshold of 0.90. This confirms that our method effectively enforces a near-perfect load balance across experts.

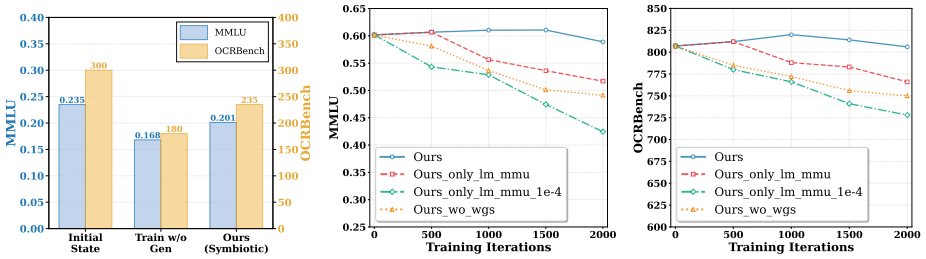
Crucially, this structural efficiency translates directly into better optimization. Figure 5b Text-to-image Training Loss reveals a turning point around 1,000 iterations, where our method surpasses the baseline MoE in convergence speed and consistently maintains a lower loss thereafter. This provides strong empirical evidence that *a compact set of 32 specialized generative experts, supported by a shared expert, is more effective than a massive pool of 128 entangled generalists plus one shared expert.* This finding underscores that for multimodal MoE, architectural specialization via the allocation of specific experts to specific modalities is more critical than simply scaling the total number of available experts.

4.3 Ablation Studies

Architectural Design Analysis. We first investigate the impact of expert partitioning and shared experts at iteration zero on knowledge retention (Table 2).

Expert Disentanglement Strategy. We first hypothesize a granular *Tripartite Split* (separating Text, ViT, and VAE). However, this configuration triggers a catastrophic collapse (MMLU $0.69 \rightarrow 0.24$), revealing a critical functional overlap between Text and ViT experts in the original VLM. Forcibly separating them severs essential semantic pathways. Consequently, we adopt a *Bimodal Split*, consolidating Text and ViT into a unified Understanding Group. By evaluating different assignment ratios, we identify the 96/32 split as the optimal configuration. This setting maximizes understanding retention (0.60) while reserving sufficient plasticity (32 experts) for the generative task.

Necessity of Shared Expert Bridge. Table 2 (row 2) further validates the role of shared experts. Removing them from the original VLM precipitates a sharp performance decline ($0.69 \rightarrow 0.46$). This confirms that shared experts act as a non-negotiable semantic anchor, ensuring feature alignment across specialized expert groups.



(a) Shared Expert Probing **(b)** MMLU Dynamics **(c)** OCRBench Dynamics

Fig. 6: Evidencing Generative Synergy. (a) We probe the isolated performance of Shared Experts. Compared to the baseline trained without generation (*Train w/o Gen*), our Symbiotic method significantly boosts the semantic density of shared experts. (b) & (c) The training dynamics show that while the control baseline degrades, Symbiotic-MoE effectively reverses the forgetting trend via generative regularization.

Optimization Dynamics Analysis. We decouple the effects of two critical strategies designed to navigate the stability-plasticity dilemma:

Differential Learning Rates. We identify a critical optimization mismatch: generative learning demands aggressive updates, whereas the pre-trained backbone requires conservative fine-tuning. Crucially, as shown by the green dashed line (`Ours_only_lm_mmu_1e-4`) in Fig. 6(b–c), training *solely* on understanding tasks at $1e^{-4}$ triggers immediate collapse even without generative interference. This exposes a fundamental insight: catastrophic forgetting here is driven less by task conflict and more by the backbone's intrinsic intolerance to high-magnitude updates. Consequently, we enforce a hierarchical strategy, assigning $1e^{-4}$ to generative experts for plasticity, while restricting shared/understanding experts to a conservative $1e^{-6}$ to respect their stability constraints.

Warmup Gradient Shielding. Generative modules trained from scratch inherently emit high-variance gradients during early optimization. Directly exposing shared experts, the model's semantic anchor, to this volatility causes an "initial shock", washing out pre-trained representations before generative features become semantically meaningful. To mitigate this, we explicitly detach the generative gradient flow to shared experts during warmup. As evidenced by the orange dashed line (`Ours_wo_wgs`) in Fig. 6(b–c), removing this shield results in a precipitous decline in MMLU and OCRBench during the first 500 iterations, confirming that early stage isolation is vital for stability.

4.4 Analysis of Synergy

While standard MoE training suffers from interference, Symbiotic-MoE orchestrates a constructive interference, synergy. To rigorously verify this, we deconstruct the interaction between modalities through three analytical lenses: component probing, subtractive ablation, and optimization dynamics.

Component Isolation of Shared Experts. To pinpoint the architectural locus of cross-modal synergy, we probe the *Shared Experts* in isolation. By applying

a hard mask to all modality-specific routed experts, we enforce zero-shot inference using only the shared expert parameters. As visualized in Fig. 6(a), the baseline (`Train w/o Gen`) optimized exclusively on understanding tasks ($\lambda_{img} = 0$) suffers from representational degradation in the shared module. In contrast, the shared experts within Symbiotic-MoE demonstrate a substantial performance resurgence, consistently outperforming the understanding-only baseline on both MMLU and OCRBench by a clear margin. This empirical evidence validates that generative gradients do not overwrite linguistic priors, rather, they function as a dense semantic compressor, forcing the shared parameters to encapsulate highly generalized representations.

Generative Regularization: Does Painting Help Seeing? We isolate the impact of generative training via a counter-factual baseline trained *solely* on understanding tasks. As visualized in Fig. 6b, the baseline (red dashed line) suffers from severe knowledge decay ($0.60 \rightarrow 0.52$ on MMLU) driven by distribution shift. Crucially, incorporating generation arrests this decline. Our Symbiotic-MoE (blue solid line) not only stabilizes general reasoning but significantly boosts fine-grained perception, propelling OCRBench to a peak of 820 (*vs*. 790). We attribute this to the unique nature of generative objectives: unlike discriminative tasks that often allow for semantic shortcuts, pixel-level reconstruction compels the shared experts to capture precise spatial relationships. This strictly constrains the representation space, effectively acting as a regularizer that prevents overfitting to textual patterns and reciprocally sharpens visual perception.

Acceleration in Optimization. Finally, we examine if strong understanding capabilities benefit generation. Figure 5b illustrates that the T2I convergence of Symbiotic-MoE is significantly faster and deeper than the Standard MoE and MoT baselines. This bi-directional flexibility aligns the semantic space with the generative manifold, turning potential conflicts into a driver for comprehensive capability emergence.

5 Conclusion

In this work, we presented Symbiotic-MoE, a unified pre-training framework that successfully reconciles the long-standing catastrophic forgetting in extending VLMs with generative capabilities. By moving beyond the structural isolation typical of prior arts like MoT, we demonstrated that task interference is not an intrinsic limitation of unified architectures, but rather an optimization challenge resolvable through modality-aware disentanglement and progressive training dynamics. Most significantly, our empirical results challenge the conventional zero-sum view of multimodal training. We show that when the routing landscape is carefully orchestrated, the fine-grained visual semantics acquired from generation can retroactively refine the model's understanding capability, rather than eroding them. We hope this work serves as a blueprint for future native omni-modal foundation models, establishing that the path to true general intelligence lies not in the segregation of perception and creation, but in their symbiotic integration within a single, efficient parameter space.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Chen, J., Tang, Q., Lu, Q., Fang, S.: Mol for llms: Dual-loss optimization to enhance domain expertise while preserving general capabilities. arXiv preprint arXiv:2505.12043 (2025)
3. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
4. Chen, Y., Wang, S., Zhang, Y., Lin, Z., Zhang, H., Sun, W., Ding, T., Sun, R.: Mofo: Momentum-filtered optimizer for mitigating forgetting in llm fine-tuning. arXiv preprint arXiv:2407.20999 (2024)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
6. Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1280–1297 (2024)
7. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
8. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
9. Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., et al.: Glam: Efficient scaling of language models with mixture-of-experts. In: International conference on machine learning. pp. 5547–5569. PMLR (2022)
10. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
11. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
12. Gu, H., Li, W., Li, L., Zhu, Q., Lee, M., Sun, S., Xue, W., Guo, Y.: Delta decompression for moe-based llms compression. arXiv preprint arXiv:2502.17298 (2025)
13. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
14. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
17. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
18. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
19. Koh, J.Y., Fried, D., Salakhutdinov, R.R.: Generating images with multimodal language models. Advances in Neural Information Processing Systems **36**, 21487–21506 (2023)
20. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
21. Liang, W., YU, L., Luo, L., Iyer, S., Dong, N., Zhou, C., Ghosh, G., Lewis, M., tau Yih, W., Zettlemoyer, L., Lin, X.V.: Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=Nu6N69i8SB
22. Lin, B., Tang, Z., Ye, Y., Huang, J., Zhang, J., Pang, Y., Jin, P., Ning, M., Luo, J., Yuan, L.: Moe-llava: Mixture of experts for large vision-language models. IEEE Transactions on Multimedia (2026)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
24. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
26. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
27. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
28. Luo, G., Yang, X., Dou, W., Wang, Z., Liu, J., Dai, J., Qiao, Y., Zhu, X.: Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24960–24971 (2025)
29. Lv, A., Ma, J., Ma, Y., Qiao, S.: Coupling experts and routers in mixture-of-experts via an auxiliary loss. arXiv preprint arXiv:2512.23447 (2025)
30. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)
31. McKinzie, B., Gan, Z., Fauconnier, J.P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Belyi, A., et al.: Mm1: methods, analysis and insights from multimodal llm pre-training. In: European Conference on Computer Vision. pp. 304–323. Springer (2024)
32. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural networks **113**, 54–71 (2019)

33. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)
34. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
35. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 (2023)
36. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems **36**, 16083–16099 (2023)
37. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
38. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
39. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
40. Wang, D., Wang, S., Li, Z., Wang, Y., Li, Y., Tang, D., Shen, X., Huang, X., Wei, Z.: Moiie: Mixture of intra-and inter-modality experts for large vision language models. arXiv preprint arXiv:2508.09779 (2025)
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
42. Wang, X., Zhang, Z., Zhang, H., Lin, Z., Zhou, Y., Liu, Q., Zhang, S., Li, Y., Liu, S., Zheng, H., et al.: Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. arXiv preprint arXiv:2511.20520 (2025)
43. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
44. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
45. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
46. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. In: Forty-first International Conference on Machine Learning (2024)
47. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
48. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13294–13304 (2025)
49. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)

50. Xu, Y., Yan, H., Cao, J., Cheng, Y., Hang, T., He, R., Yin, Z., Zhang, S., Zhang, Y., Li, J., et al.: Tag-moe: Task-aware gating for unified generative mixture-of-experts. arXiv preprint arXiv:2601.08881 (2026)
51. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
52. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
53. Zhang, H., Gao, M., Gan, Z., Dufter, P., Wenzel, N., Huang, F., Shah, D., Du, X., Zhang, B., Li, Y., et al.: Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. arXiv preprint arXiv:2409.20566 (2024)
54. Zhang, L., Xie, Y., Fang, F., Dong, F., Liu, R., Cao, Y.: Metagdpo: Alleviating catastrophic forgetting with metacognitive knowledge through group direct preference optimization. arXiv preprint arXiv:2511.12113 (2025)
55. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
56. Zhang, Z., Dong, Q., Zhang, Q., Zhao, J., Zhou, E., Xi, Z., Jin, S., Fan, X., Zhou, Y., Fu, Y., et al.: Reinforcement fine-tuning enables mllms learning novel tasks stably. arXiv e-prints pp. arXiv–2506 (2025)
57. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)
58. Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., Fedus, W.: St-moe: Designing stable and transferable sparse expert models. arXiv preprint arXiv:2202.08906 (2022)