Preventing Expert Collapse in MoE-dVLMs via Modality-Wise Norm Alignment

Zongkai Liu[1,2,3], Zhen Cao[3], Hui Zhang[3], Chao Yu[1,†], Liqiang Niu[3], and Fandong Meng[3,†]

[1] Sun Yat-sen University, Guangzhou, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] Weixin AI, Tencent, Shanghai, China
liuzk@mail2.sysu.edu.cn, yuchao3@mail.sysu.edu.cn, fandongmeng@tencent.com

Abstract. Sparse Mixture-of-Experts (MoE) offers an attractive path to scaling discrete diffusion vision–language models (dVLMs), but we find that naïvely combining MoE with multimodal diffusion training can fail catastrophically. Specifically, MoE-dVLMs suffer from *expert selection collapse* on visual tokens, where most image tokens are routed to only a few experts and overall multimodal performance degrades sharply. We trace this failure to an upstream geometric mismatch at the modality boundary rather than router optimization. Visual embeddings produced by deep Pre-Norm vision encoders have much larger norms than text embeddings; when injected into a Pre-Norm language tower, the residual dynamics cause *attention drowning* and *representational inertia*, yielding highly concentrated, low-rank visual representations that are difficult for the router to discriminate. To resolve this issue, we propose **TowerAlign**, a lightweight post-projector modality-wise norm alignment that globally rescales visual tokens to match text norm statistics while preserving their relative structure. TowerAlign stabilizes MoE routing and improves multimodal accuracy with negligible overhead. On a practical MoE-dVLM built from LLaDA-MoE-7B-A1B with a SigLIP2-SO400M-Patch14-384 vision encoder, TowerAlign delivers consistent gains across diverse benchmarks and largely closes the gap to dense diffusion baselines under matched data budgets.

Keywords: Discrete Diffusion Vision-Language Models · Mixture-of-Experts · Modality-Wise Norm Alignment

1 Introduction

Mixture-of-Experts (MoE) is a dominant approach to scaling language models, boosting parameter capacity while keeping inference cost low via sparse activation [5, 6]. In parallel, discrete masked diffusion language models (dLLMs) have

† Corresponding author.

rapidly closed the performance gap to autoregressive LLMs while offering bidirectional context modeling and flexible decoding [3,21,33]. These advances naturally motivate an appealing direction: building diffusion vision-language models (dVLMs) by augmenting diffusion backbones with sparse MoE architectures.

However, when we bring MoE into dVLMs (MoE-dVLMs), a critical failure mode emerges. Training recipes that work well for dense dVLMs can degrade drastically in MoE variants: we observe *expert selection collapse* where visual tokens are routed to a tiny subset of experts, leaving most experts under-utilized and harming downstream performance. Importantly, this collapse persists even with strong load-balancing regularization, indicating that the root cause is not simply an ill-optimized router.

We trace the failure to a geometric incompatibility at the tower interface. Visual encoders (deep Pre-Norm transformers) often output embeddings with substantially larger L2 norms than textual embeddings. When these high-norm visual tokens are injected into a language-model tower calibrated for low-norm text, the Pre-Norm residual dynamics induce *attention drowning*: the self-attention update becomes negligible compared to the unnormalized residual, sharply reducing the effective contribution of token interactions. This suppression yields *representational inertia*—visual tokens change direction only marginally across layers—leading to high cosine similarity and low effective rank in the visual embedding space (Section 4). Under such geometric concentration, the router cannot reliably discriminate tokens: top-k expert selections become effectively identical across visual positions, and a Matthew-effect feedback loop amplifies the bias as over-selected experts receive larger updates and grow stronger.

Based on this diagnosis, we propose **TowerAlign**, a simple modality-wise norm alignment applied immediately after the visual projector. TowerAlign rescales projected visual tokens to match the norm statistics of textual embeddings, restoring symmetric update dynamics in the Pre-Norm tower while preserving the relative structure among visual tokens. Crucially, this intervention is *lightweight* and statistics-driven, incurs negligible additional training cost, and directly targets the upstream geometry that causes collapse rather than attempting to fix routing downstream.

We instantiate TowerAlign in a practical MoE-dVLM by extending LLaDA-MoE-7B-A1B [33] with a SigLIP2-SO400M-Patch14-384 [24] vision encoder and an 2-layer MLP projector. Across multimodal benchmarks, TowerAlign yields consistent gains and largely closes the gap to dense baselines. Our findings suggest that geometric alignment at the modality boundary is both necessary and sufficient for scalable MoE-based dVLMs.

Our contributions are threefold: (1) we identify expert selection collapse on visual tokens as a key failure mode in MoE-dVLMs pretraining and show it stems from modality-wise norm discrepancy and attention drowning; (2) we propose TowerAlign, a lightweight modality-wise scale alignment that restores attention efficacy and router distinguishability; and (3) we establish an effective training recipe for MoE-dVLMs with competitive empirical results.

2 Related Work

2.1 Vision-Language Models and the Modality Gap Challenge

The architectural evolution of VLMs has undergone a significant paradigm shift in recent years. Early approaches such as Flamingo [1] and BLIP-2 [11] relied on cross-attention mechanisms for modality fusion. The contemporary dominant paradigm, popularized by LLaVA [14], adopts an **MLP projection module** to map visual tokens directly into the LLM's embedding space, enabling joint processing of visual and textual inputs through standard transformer architectures. This streamlined design has been widely adopted by subsequent VLMs such as Qwen-VL series [2, 25] and InternVL series [4].

Recent work has identified the **modality gap** as a critical phenomenon in multimodal architectures, where visual and textual representations occupy systematically offset regions in the joint embedding space despite expressing identical semantics [9, 22, 29]. Analyses reveal that this gap comprises not merely centroid shifts but complex anisotropic residuals that hinder cross-modal interchangeability [29]. Several recent studies further address modality imbalance in dense multimodal models: mean-shift correction has been explored for CLIP-style retrieval gaps [8], self-distillation has been used to mitigate VLM reading errors [23], and post-projector norm rescaling has been shown to reduce ViT–LLM mismatch and representational inertia in dense VLMs [10]. These works establish that cross-modal geometry matters, but they do not study sparse expert routing. Our contribution is therefore not merely applying rescaling, but identifying a MoE-specific failure mode: severe visual–text norm discrepancy (e.g., 167× magnitude differences) geometrically concentrates visual tokens, making explicit modality alignment critical for avoiding **expert selection collapse**.

2.2 Discrete Masked Diffusion (Vision-)Language Models

dLLMs have emerged as a compelling alternative to autoregressive generation, offering potential for non-monotonic reasoning and parallel decoding. Recent advances have scaled dLLMs to unprecedented sizes. Notably, LLaDA [21] scaled masked diffusion models to 8B parameters, demonstrating competitive performance with LLaMA3-8B-Instruct. Subsequently, LLaDA2.0 [3] has further pushed this frontier to **100B parameters** through systematic conversion from autoregressive models, introducing Warmup-Stable-Decay training strategies and block diffusion mechanisms.

In the multimodal domain, dVLMs have emerged as a promising alternative to autoregressive approaches. **LaViDa** [12] concurrently explored this paradigm, introducing two distinct variants: **LaViDa-L**, which employs LLaDA-8B [21] as the diffusion backbone, and **LaViDa-D**, which utilizes Dream-7B [27]. These variants demonstrated competitive performance against autoregressive VLMs across multimodal benchmarks while offering unique advantages inherent to diffusion models, including flexible speed-quality tradeoffs and bidirectional reasoning capabilities.

In parallel, **LLaDA-V** [28] advanced the paradigm through **large-scale visual instruction tuning**, employing extensive multimodal datasets and multistage training strategies (alignment, visual instruction tuning, and reasoning enhancement) to achieve performance competitive with autoregressive VLMs. Notably, LLaDA-V achieved comparable performance to autoregression VLMs like LLaMA3-V [28] on multidisciplinary knowledge and mathematical reasoning benchmarks despite using a weaker language tower, demonstrating that purely diffusion-based VLMs can match autoregressive performance on complex multimodal tasks through sufficient data scaling and proper training recipes.

Despite these advances in dVLMs, the integration of discrete diffusion with sparse MoE architectures for multimodal understanding remains unexplored. This leaves open whether training recipes that are effective for dense dVLMs remain valid once sparse routing is introduced, and what new failure modes emerge from the interaction between multimodal representation geometry and expert selection.

2.3 Mixture-of-Experts for Diffusion Models

MoE has become a standard approach for scaling language models via sparse activation [5,6]. To prevent degenerate expert usage, prior MoE systems commonly employ routing-side regularization, such as auxiliary load-balancing objectives that encourage balanced expert selection across a batch [6]. These methods address imbalance caused by router optimization or expert competition. Our setting differs in cause: visual tokens can become geometrically indistinguishable before reaching the router, so their soft routing probabilities may remain high-entropy while greedy top-k selection still assigns them to the same few experts. In this regime, stronger load balancing provides little gradient signal for escaping collapse.

Recent efforts have begun to bring MoE into dLLMs: **LLaDA-MoE** [33] demonstrates that sparse experts can match dense dLLMs, and OpenMoE2 [20] further studies scaling behavior. In contrast, extending MoE-based dLLM backbones to *multimodal* settings remains unexplored. In this work, we show that multimodal inputs introduce modality-specific representation geometry that interacts with routing in non-trivial ways: a severe norm mismatch between visual and textual embeddings geometrically concentrates visual tokens and triggers *expert selection collapse* (Section 4), rendering routing-side regularization ineffective. We further demonstrate that aligning modality-wise norm statistics at the tower interface (TowerAlign) provides a simple, robust remedy that restores stable and effective MoE training.

3 Preliminaries

We briefly review the architectural components central to our analysis: dVLMs, MoE mechanisms, and the Pre-Norm transformer design. Understanding their interactions is essential for diagnosing expert selection collapse in MoE-dVLMs.

3.1 Discrete Diffusion Vision-Language Models

Similarly to autoregression VLM, a standard dVLM comprises three components: (1) a vision encoder (the "vision tower") that extracts a sequence of visual features $\mathbf{z}_v \in \mathbb{R}^{L_v \times d_v}$ from an input image $\mathbf{x}_v$; (2) an MLP projector $P(\cdot)$ that maps visual features into the language embedding space, yielding $\mathbf{h}_v = P(\mathbf{z}_v) \in \mathbb{R}^{L_v \times d}$; and (3) a dLLM (the "language tower") that processes an interleaved multimodal sequence. Concretely, we conceptually split the textual stream into a *prompt* $\mathbf{p}$ and a *response* $\mathbf{y}$, and write $\mathbf{h}^{(0)} = [\mathbf{h}_v; \mathbf{h}_p; \mathbf{h}_y] \in \mathbb{R}^{L \times d}$, where $\mathbf{h}_p \in \mathbb{R}^{L_p \times d}$ and $\mathbf{h}_y \in \mathbb{R}^{L_y \times d}$ are embeddings of the prompt and response tokens, respectively, and $L = L_v + L_p + L_y$.

Masked diffusion objective. Let $\mathbf{y} = (y^1, y^2, \ldots, y^{L_y})$ denote a ground-truth response sequence of length L_y from a vocabulary of size K. The corruption process samples a noise level $\tau \sim \mathcal{U}[0,1]$, then independently replaces each response token y^i with a mask token $\mathbf{M}$ with probability τ. The forward process is

$$q(\mathbf{y}_\tau|\tau,\mathbf{y}) = \prod_{i=1}^{L_y} q(y_\tau^i|\tau,y^i), \qquad q(y_\tau^i \mid \tau, y^i) = \begin{cases} 1-\tau, & y_\tau^i = y^i, \\ \tau, & y_\tau^i = \mathbf{M}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The diffusion transformer consumes the concatenated sequence $[\mathbf{h}_v; \mathbf{h}_p; \mathbf{h}_{y,\tau}]$ of length $L = L_v + L_p + L_y$, where $\mathbf{h}_{y,\tau}$ are embeddings of the corrupted response tokens $\mathbf{y}_\tau$ (the prompt embeddings $\mathbf{h}_p$ are kept intact). We train a predictor $p_\theta(\cdot \mid \mathbf{x}_v, \mathbf{p}, \mathbf{y}_\tau)$ to reconstruct the clean response tokens. Following LLaDA, we optimize

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{p},\mathbf{y})\sim p_{\mathrm{data}}} \mathbb{E}_{\tau\sim\mathcal{U}[0,1]} \mathbb{E}_{\mathbf{y}_\tau\sim q(\mathbf{y}_\tau|\tau,\mathbf{y})} \Bigg[\frac{1}{\tau} \sum_{i=1}^{L_y} \mathbf{1}[y_\tau^i = \mathbf{M}] \log p_\theta(y^i \mid \mathbf{x}_v, \mathbf{p}, \mathbf{y}_\tau) \Bigg]. \tag{2}$$

During inference, generation proceeds via iterative denoising from a fully masked response initialization.

3.2 The Pre-Norm Transformer Architecture

Modern language models predominantly adopt the Pre-Norm design, which normalizes the input to each sublayer while keeping the residual path unnormalized. A Pre-Norm Transformer block (attention + feed-forward) can be written as

$$\mathbf{u}^{(l)} = \mathbf{h}^{(l)} + \mathrm{SelfAttn}(\mathrm{RMSNorm}(\mathbf{h}^{(l)})), \tag{3}$$

$$\mathbf{h}^{(l+1)} = \mathbf{u}^{(l)} + \mathrm{FFN}(\mathrm{RMSNorm}(\mathbf{u}^{(l)})), \tag{4}$$

where

$$\mathrm{RMSNorm}(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\frac{1}{d}\|\mathbf{x}\|_2^2 + \epsilon}} \odot \mathbf{b}. \tag{5}$$

Here $\mathbf{b} \in \mathbb{R}^d$ is a learnable parameter and $\epsilon > 0$. Because the residual branches bypass normalization, the hidden-state norm can accumulate with depth. This asymmetry is central to our analysis: when token norms differ substantially across modalities, the effective contribution of attention/FFN updates can become highly unbalanced, setting the stage for attention drowning and downstream routing collapse.

3.3 Mixture-of-Experts in Language Models

MoE scales capacity by replacing dense feed-forward layers with a set of sparse expert networks, activating only a small subset per token. An MoE layer consists of N experts $\{E_1, \ldots, E_N\}$ and a router network $\mathrm{Router}(\cdot)$. Given a token hidden state $\mathbf{u} \in \mathbb{R}^d$, the router outputs gating probabilities $\mathbf{g} = \mathrm{softmax}(\mathrm{Router}(\mathbf{u})) \in \mathbb{R}^N$. The top-$k$ experts are selected and combined as

$$\mathrm{MoE}(\mathbf{u}) = \sum_{i \in \text{top-}k(\mathbf{g})} g_i \, E_i(\mathbf{u}). \tag{6}$$

To avoid degenerate routing, auxiliary load-balancing losses are commonly used to encourage balanced expert utilization. In this work, we show that in MoE-dVLMs, visual-token collapse can persist even under strong load-balancing, motivating an upstream geometric diagnosis.

4 Diagnosing Expert Collapse in MoE-dVLMs

4.1 Performance Degradation and Emergence of Expert Collapse

The two-stage training paradigm pioneered by LLaVA has established itself as the de facto standard for VLM training due to its elegant simplicity and practical efficacy. This paradigm comprises: (1) a **feature alignment pretraining phase** where only the projector is optimized to map visual tokens into the linguistic embedding space, followed by (2) an **end-to-end visual instruction tuning phase** involving joint optimization of all components.

We adopt this training strategy in our experimental setup: for pretraining, we employ the LCS-558K dataset [14] consisting of 558K image-text pairs; for finetuning, we utilize the LLaVA-NeXT dataset [13] containing 780K multi-turn conversational samples with interleaved image-text sequences. When applying this proven strategy to dense models such as LLaDA-8B-Instruct, we observe consistent convergence and competitive benchmark performance. However, a perplexing phenomenon emerges when transitioning to its MoE variant, LLaDA-MoE-7B-A1B: despite architectural similarity and identical training configurations (optimizer hyperparameters and data pipelines), the MoE version exhibits

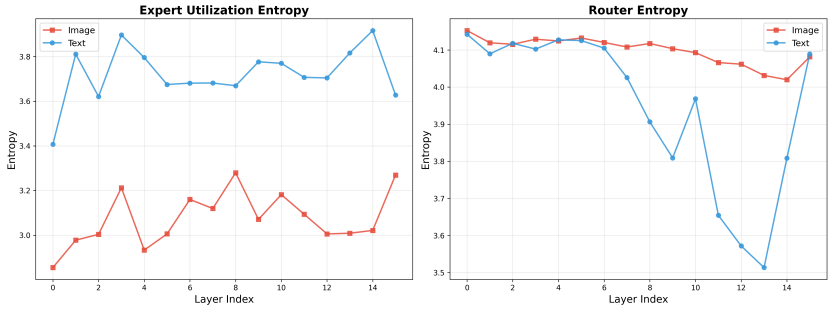

Fig. 1: Expert Utilization vs. Router Entropy. The left plot shows expert utilization entropy computed from the empirical top-k assignment frequency, while the right plot shows router entropy computed from the softmax gating distribution, for visual (Image) and textual (Text) tokens across layers.

systematic performance degradation that far exceeds the scope of standard optimization variance. Concretely, compared to the matched-budget dense diffusion baseline LLaDA-VLM (Table 2), the MoE model scores only 4.7 on ChartQA and 13.8 on DocVQA, *vs*. 57.4 and 61.2 for LLaDA-VLM.

To trace the origins of this performance chasm, we conducted a systematic dissection of the MoE training dynamics. Examination of training logs revealed a critical anomaly: the model displayed severe expert selection collapse during visual token processing, wherein a minority of experts monopolized the computational flow for the vast majority of visual tokens. Figure 1 illustrates this phenomenon through the lens of Expert Utilization Entropy, which is the entropy calculated using the frequency of expert usage: visual tokens exhibit significantly lower entropy ($\bar{\mathcal{H}}_{\text{vis}} \approx 3.08$) compared to textual tokens ($\bar{\mathcal{H}}_{\text{txt}} \approx 3.72$), indicating highly concentrated expert assignment for visual inputs.

4.2 Initial Diagnosis: Beyond Load-Balancing Losses

Initially, we hypothesized that this collapse stemmed from deficiencies in the routing mechanism itself. Consequently, we imposed aggressive regularization strategies to enforce expert equilibrium, including:

- **Load-balancing losses:** Auxiliary losses penalizing imbalance in expert utilization (coefficients scaled up to 10× baseline, i.e., 0.1);
- **Gumbel-Softmax perturbations:** Stochastic noise injection ($\tau = 2.0$) to break deterministic routing.

Strikingly, however, the skewed distribution of expert loads proved remarkably resilient to these interventions. Figure 2 compares the expert utilization entropy across layers under these two regularizers. Rather than alleviating collapse, both strategies can further *reduce* visual-token utilization entropy (in some layers even lower than the unregularized setting) and simultaneously lead to

worse downstream performance. Indeed, Table 2 shows that both interventions severely degrade accuracy (e.g., on DocVQA, Gumbel drops to 6.8 and stronger **LB loss** to 9.3). This suggests that such routing-side regularization not only fails to address expert selection collapse, but can also interfere with optimizing the primary training objective, ultimately harming overall model quality.

This observation prompted us to scrutinize the router's behavioral patterns more closely. Through visualization of the raw routing probability distributions as shown in Figure 1, we found that the router entropy for visual tokens is comparable to, and in some layers even higher than, that for textual tokens. In other words, the softmax probabilities do not exhibit the extreme skewness characteristic of a collapsed gating distribution.

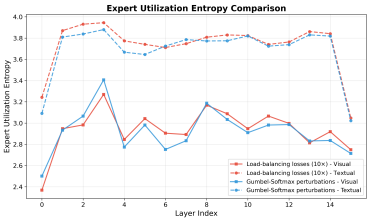

Fig. 2: Expert Utilization Entropy under Routing Regularization.

The true culprit, we identified, lies in the geometric structure of the visual embedding space: visual token representations become nearly degenerate, implying high similarity across different spatial positions. We quantify this via singular value decomposition of token-embedding matrices. Concretely, we use the per-sample visual embedding matrix $\mathbf{h}_v \in \mathbb{R}^{L_v \times d}$, and define the textual embedding matrix as $\mathbf{h}_t = [\mathbf{h}_p; \mathbf{h}_y] \in \mathbb{R}^{L_t \times d}$ with $L_t = L_p + L_y$. Using the effective rank (computed from the singular-value spectrum), we find a striking collapse: for our model with $d = 2048$, the effective rank of visual tokens is only 2.1/2048, indicating that the visual embedding distribution effectively lives on an extremely low-dimensional subspace; in contrast, textual tokens achieve 1673.3/2048, suggesting a much richer and more isotropic representation space.

This *low-rank collapse* implies that visual tokens occupy a degenerate low-dimensional manifold within the high-dimensional embedding space. Consequently, when projected onto the router's weight space, these embeddings span only a narrow cone, rendering the routing decision effectively low-dimensional regardless of the apparent entropy in probability distributions.

4.3 The Root Cause: Norm Discrepancy and Asymmetric Update Dynamics

Behind this representational homogenization, we identify a critical mediating variable: the L2 norm disparity between visual and textual embeddings. Let α_v and α_t denote the mean per-token L2 norms at the language-tower input for visual tokens $\mathbf{h}_v$ and textual tokens $\mathbf{h}_t = [\mathbf{h}_p; \mathbf{h}_y]$, respectively. Empirically, $\alpha_v \approx 60.47 \pm 108.34$ while $\alpha_t \approx 0.36 \pm 0.13$, i.e., a staggering $\sim$ **167**$\times$ gap.

Recall from our preliminaries that Pre-Norm blocks normalize only the sublayer input while leaving the residual branch unnormalized (Eq. 3), so main-path hidden-state norms can accumulate with depth. For a token index i, we denote

its modality-specific hidden state by $\mathbf{h}_{v,i}^{(l)}$ (visual) or $\mathbf{h}_{t,i}^{(l)}$ (text), and similarly $\mathbf{u}_{v,i}^{(l)}$ and $\mathbf{u}_{t,i}^{(l)}$ after the attention residual.

Crucially, the attention and (MoE) FFN sublayers are text-pretrained and RMSNormed, so their output scales are calibrated to be relatively small (on the order of text-token updates) [9]. We can thus view the per-layer update magnitudes as approximately modality-agnostic constants:

$$\|\mathrm{SelfAttn}(\mathrm{RMSNorm}(\mathbf{h}_{m,i}^{(l)}))\|_2 \approx C_{\mathrm{att}}^{(l)}, \quad \|\mathrm{FFN}(\mathrm{RMSNorm}(\mathbf{u}_{m,i}^{(l)}))\|_2 \approx C_{\mathrm{ffn}}^{(l)},$$

for $m \in \{v, t\}$. Therefore, when $\|\mathbf{h}_{v,i}^{(l)}\|_2 \approx \alpha_v \gg \alpha_t \approx \|\mathbf{h}_{t,i}^{(l)}\|_2$, adding an $\mathcal{O}(\alpha_t)$ update to an $\mathcal{O}(\alpha_v)$ residual causes the interaction information to be *drowned* by the residual path, i.e., the relative change in direction becomes negligible. The same drowning phenomenon also applies to the MoE FFN block, since it shares the same Pre-Norm residual form.

This drowning effect yields a dramatic attenuation of effective angular velocity. According to the theoretical derivation in [9], the rate of directional change for a token from modality m follows:

$$\tan(\theta_{\mathrm{eff},m,i}^{(l)}) = \frac{C^{(l)} \sin(\theta_{m,i}^{(l)})}{\|\mathbf{h}_{m,i}^{(l)}\|_2 + C^{(l)} \cos(\theta_{m,i}^{(l)})},$$

where $\theta_{m,i}^{(l)}$ represents the expected angle between the update direction and the current hidden state. When $\alpha_v \gg \alpha_t$, visual tokens exhibit much smaller effective angular velocities, e.g., $\theta_{\mathrm{eff}}, v \approx \arctan(C^{(l)}/\alpha_v) \approx 1^\circ$. This asymmetry results in **representational inertia** [9]: visual tokens traverse the embedding space along nearly straight trajectories with minimal directional exploration, while textual tokens undergo substantial angular dispersion.

In the context of MoE architectures, this representational inertia is further amplified by the routing mechanism. When highly homogenized visual tokens (due to insufficient directional updates) reach the router, even superficially different logit distributions produce nearly identical top-k selections because similar direction vectors project similarly onto expert weight spaces. This explains why conventional load-balancing losses fail: **the root cause lies not in router optimization, but in the upstream representation space lacking sufficient directional diversity to support meaningful routing discrimination.**

The foregoing analysis receives rigorous support from our theoretical framework, which formally establishes that high-norm embeddings inevitably converge to a low-dimensional manifold, causing the observed similarity explosion:

Theorem 1 (High-Norm Induced Representation Collapse). *Let* $\{\mathbf{h}_{v,i}^{(l)}\}_{i=1}^{L_v} \subset \mathbb{R}^d \setminus \{0\}$ *denote the* L_v *visual-token embeddings at layer* l *of a depth-*ν *Transformer. Assume:*

1. *(Bounded Norm): There exists* $\alpha_v > 1$ *such that for all* $i \in \{1, \ldots, L_v\}$ *and* $l \in \{0, \ldots, \nu - 1\}$, $\|\mathbf{h}_{v,i}^{(l)}\|_2 = \alpha_v$.

2. *(Gradient Structure):* $\nabla_{\mathbf{h}_{v,i}^{(l)}} \mathcal{L} = c_i^{(l)} \mathbf{v} + \boldsymbol{\epsilon}_i^{(l)}$, *where* $\mathbf{v} \in \mathbb{S}^{d-1}$ *is a fixed principal direction,* $c_i^{(l)} \in [c_{\min}, c_{\max}] \subset (0, \infty)$, *and* $\|\boldsymbol{\epsilon}_i^{(l)}\|_2 \leq \delta \ll 1$.
3. *(Update Rule): Under a Pre-Norm update with learning rate* $\eta > 0$, $\mathbf{h}_{v,i}^{(l+1)} = RMSNorm(\mathbf{h}_{v,i}^{(l)} - \eta \nabla_{\mathbf{h}_{v,i}^{(l)}} \mathcal{L})$, *where we use the simplified form* $RMSNorm(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ *for analysis.*
4. *(Effective Angle Constraint): There exists* $C > 0$ *such that* $\sin(\theta_{v,i}^{(l)}) \leq \frac{C}{\alpha_v}$, *where* $\theta_{v,i}^{(l)}$ *is the angle between* $-\nabla_{\mathbf{h}_{v,i}^{(l)}} \mathcal{L}$ *and* $\mathbf{h}_{v,i}^{(l)}$.

Define direction vectors $\hat{\mathbf{h}}_{v,i}^{(l)} = \mathbf{h}_{v,i}^{(l)}/\|\mathbf{h}_{v,i}^{(l)}\|_2$. *Then for all* $i, j \in \{1, \ldots, L_v\}$,

$$1 - \langle \hat{\mathbf{h}}_{v,i}^{(\nu)}, \hat{\mathbf{h}}_{v,j}^{(\nu)} \rangle \leq \frac{2C^2\nu}{\alpha_v^2} + \frac{4\delta\nu}{\alpha_v} + \Big(1 - \langle \hat{\mathbf{h}}_{v,i}^{(0)}, \hat{\mathbf{h}}_{v,j}^{(0)} \rangle\Big).$$

In particular, if the initial directions are roughly uniform so that $\langle \hat{\mathbf{h}}_{v,i}^{(0)}, \hat{\mathbf{h}}_{v,j}^{(0)} \rangle \approx 0$, *then*

$$\langle \hat{\mathbf{h}}_{v,i}^{(\nu)}, \hat{\mathbf{h}}_{v,j}^{(\nu)} \rangle \gtrsim 1 - \frac{2C^2\nu}{\alpha_v^2} - \frac{4\delta\nu}{\alpha_v}.$$

Proof. The rationality of assumptions and the complete proof can be found in Appendix A.

5 TowerAlign: Modality-Wise Norm Alignment

The diagnosis presented in the preceding section reveals that the crux of MoE dysfunction in dVLMs lies not in the routing mechanism itself, but in the profound geometric asymmetry between visual and textual modalities. Specifically, the approximately 167-fold disparity in L2 norms induces severe representational inertia within the Pre-Norm architecture, effectively freezing the angular evolution of visual representations across layers. This inertia collapses the diversity of the visual embedding manifold, rendering the MoE router incapable of generating meaningfully differentiated assignment decisions.

To resolve this pathology, we propose TowerAlign, a lightweight, post-projector statistical alignment strategy that restores symmetric update dynamics without compromising the intrinsic semantic hierarchy of visual tokens. The fundamental premise of TowerAlign is elegantly simple: we globally rescale the radial magnitude of visual embeddings to match that of textual embeddings (target mean $\mu = 0.36$), while meticulously preserving the relative geometric relationships among visual tokens.

5.1 Global Affine Compression vs. RMSNorm Baseline

Global Affine Compression (GAC). The primary implementation of TowerAlign employs a global affine transformation applied immediately following the visual

projector. Let $\mathbf{H}_v^{(0)} \in \mathbb{R}^{B\times L_v\times d}$ denote the projected visual embeddings at the language-tower input, and let $\mathbf{h}_{v,b,i}^{(0)} \in \mathbb{R}^d$ denote the embedding of the i-th visual token in the b-th sample. We compute a single scalar scaling factor by averaging token norms over *both* the batch and token dimensions:

$$\tilde{\mathbf{h}}_{v,b,i}^{(0)} = \lambda \cdot \mathbf{h}_{v,b,i}^{(0)}, \quad \lambda = \frac{\mu_{\text{target}}}{\frac{1}{BL_v}\sum_{b=1}^{B}\sum_{i=1}^{L_v}\|\mathbf{h}_{v,b,i}^{(0)}\|_2}.$$

This operation constitutes a structure-preserving radial compression. Critically, because every visual token is multiplied by the identical scalar λ, the relative norm ratios between any pair of tokens remain invariant. For any two visual tokens $\mathbf{h}_{v,i}^{(0)}$ and $\mathbf{h}_{v,j}^{(0)}$, we have $\|\tilde{\mathbf{h}}_{v,i}^{(0)}\|_2/\|\tilde{\mathbf{h}}_{v,j}^{(0)}\|_2 = \|\mathbf{h}_{v,i}^{(0)}\|_2/\|\mathbf{h}_{v,j}^{(0)}\|_2$. Thus, while the absolute magnitude of the visual distribution is compressed to align with the textual embedding space (restoring comparable effective angular velocities for residual updates), the internal angular topology of the visual manifold remains intact. High-norm tokens retain their relative prominence over low-norm tokens, preserving the latent semantic signaling encoded in these magnitude disparities.

RMSNorm Baseline. As a comparative baseline, we evaluate a per-token normalization strategy based on RMSNorm:

$$\tilde{\mathbf{h}}_{v,b,i}^{(0)} = \text{RMSNorm}(\mathbf{h}_{v,b,i}^{(0)}) \cdot \frac{\mu_{\text{target}}}{\sqrt{d}}.$$

This approach enforces strict second-moment uniformity across all tokens, theoretically eliminating norm-based disparities within the visual sequence. However, as our empirical results demonstrate in Table 2, this aggressive homogenization proves detrimental to model performance, offering critical insights into the information geometry of vision transformers.

The superiority of Global Affine Compression over RMSNorm illuminates a subtle yet crucial aspect of Vision Transformer representations: the phenomenon of ViT Sink tokens. Recent analyses of deep ViTs reveal the existence of a small subset of tokens (typically 3–5 per image, often corresponding to semantically salient regions or background patches) that exhibit exceptionally high L2 norms (frequently exceeding 100) and concentrated activation patterns in specific hidden dimensions [16]. These Sink tokens function as information aggregation hubs, encoding high-level global semantics such as scene categories, overall illumination, and compositional structure [16].

Conversely, the majority of Non-Sink tokens exhibit more moderate norms (typically around 60 in our observations) and encode localized, fine-grained visual features. Crucially, the ratio between Sink and Non-Sink token norms serves as an implicit signaling mechanism: the elevated magnitude of Sink tokens indicates their role as carriers of global contextual information, allowing downstream layers to differentially attend to these information-dense representations [16].

RMSNorm disrupts this signaling structure. By individually normalizing each token to unit RMS followed by uniform scaling, RMSNorm forces both Sink and

Non-Sink tokens into an identical magnitude regime (μ_{target}), effectively erasing the distributional cues that distinguish global semantics from local details, and thus reducing the performance.

In contrast, TowerAlign's global compression preserves these essential relative proportions, ensuring that Sink tokens maintain their relative salience in the dot-product calculations underlying MoE routing. This preservation allows the router to effectively leverage global semantic cues for expert assignment, directing computationally intensive global reasoning to appropriate experts while delegating local feature processing to others.

6 Experiments

6.1 Experimental Setup

Model Configuration. We instantiate our MoE-dVLM based on LLaDA-MoE-7B-A1B [33], a discrete diffusion language model with 7B total parameters and 1B active parameters per token. The architecture comprises 64 experts with top-8 routing. For the visual tower, we employ SigLIP2-SO400M-Patch14-384 [24] as the vision encoder, followed by a two-layer MLP projector with GELU activation. The experiment was conducted with a maximum sequence length of 8192 tokens.

Training Protocol. We strictly adhere to the LLaVA two-stage training paradigm to ensure fair comparison with dense baselines:

- **Stage 1 (Feature Alignment):** We train only the projector for 1 epoch on the LCS-558K dataset (558K image-text pairs), using a learning rate of 1×10^{-3} with cosine decay and warmup ratio 0.03.
- **Stage 2 (Visual Instruction Tuning):** We perform end-to-end fine-tuning on the single-image data of LLaVA-NeXT dataset (780k multi-turn conversations), updating the full model including the vision tower, projector, and language tower. The learning rate is set to 1×10^{-5} for the language model and 2×10^{-6} for the vision encoder, with the same scheduler configuration.

Evaluation Benchmarks. We evaluate on comprehensive multimodal benchmarks spanning: (1) **chart and document understanding** (ChartQA [17], DocVQA [19], InfoVQA [18]); (2) **real-world scene understanding** (RealWorldQA [26]); (3) **scientific diagram reasoning** (AI2D [7]); and (4) **multidisciplinary knowledge and reasoning** (MMBench [15], MMMU [30] and MMMU-Pro [31]). All benchmarks are evaluated with the `lmms-eval` [32] toolkit.

6.2 Main Results

Table 1 reports the performance of our MoE and dense baselines, together with representative diffusion-based and autoregressive VLMs, on our evaluation suite. Specifically, LaViDa [12] uses LLaDA-Instruct-8B [21] (LaViDa-L) and Dream-7B [27] (LaViDa-D) as language backbones; LLaDA-V [28] is built on LLaDA-Instruct-8B; and LLaVA-1.6 [13] is an autoregressive VLM.

Table 1: Main results on multimodal benchmarks. We report the official score for each benchmark.

Benchmark	**TowerAlign**	**LLaDA-VLM**	**LaViDa-L**	**LaViDa-D**	**LLaDA-V**	**LLaVA-1.6**
#Params	7B-A1B	8B	8B	7B	8B	7B
#Pretrain	0.6M	0.6M	0.6M	0.6M	0.6M	0.6M
#SFT	0.8M	0.8M	1.0M	1.0M	16.0M	0.7M
Type	Diffusion	Diffusion	Diffusion	Diffusion	Diffusion	Autoregressive
ChartQA	59.0	57.4	64.6	61.0	78.3	54.8
DocVQA (val)	67.7	61.2	59.0	56.1	83.9	74.4
InfoVQA (val)	25.8	25.5	34.2	36.2	66.3	37.1
RealWorldQA	58.9	57.4	–	–	63.2	–
MMStar	46.7	46.0	–	–	60.1	–
AI2D	70.5	70.2	70.0	69.0	77.8	66.6
MMBench (EN)	73.3	71.9	70.5	73.8	82.9	54.6
MMMU (val)	40.3	40.1	43.3	42.6	48.6	35.1
MMMU-Pro (standard)	25.2	25.4	–	–	35.2	–
MMMU-Pro (vision)	15.1	15.7	–	–	18.6	–

Overall, TowerAlign demonstrates that a sparse MoE-dVLM (7B-A1B, i.e., only 1B active parameters per token) can match or even rival dense 7B–8B baselines across diverse multimodal benchmarks. With matched data budgets to the dense diffusion baseline (LLaDA-VLM; 0.6M pretraining + 0.8M SFT), TowerAlign achieves consistent gains on chart/document and general understanding tasks such as ChartQA (59.0 vs. 57.4), DocVQA (67.7 vs. 61.2), RealWorldQA (58.9 vs. 57.4), and MMBench (73.3 vs. 71.9), while remaining comparable on AI2D (70.5 vs. 70.2) and MMMU (40.3 vs. 40.1). The only notable regressions are on MMMU-Pro (standard: 25.2 vs. 25.4; vision: 15.1 vs. 15.7).

Compared with the dense 7B autoregressive model (LLaVA-1.6), TowerAlign is substantially stronger on reasoning-heavy benchmarks such as AI2D (70.5 vs. 66.6), MMBench (73.3 vs. 54.6), and MMMU (40.3 vs. 35.1), despite using diffusion decoding and sparse activation. Meanwhile, a sizable gap remains to LLaDA-V, which benefits from a much larger instruction-tuning scale (16.0M SFT). Together, these results support our core claim: **aligning modality-wise norm statistics mitigates MoE expert collapse and enables strong multimodal performance with sparse activation and negligible overhead.**

6.3 Ablation Studies

We conduct ablation studies to isolate the contributions of (i) modality-wise norm alignment, (ii) the specific alignment operator, (iii) routing-side regularization baselines, and (iv) multi-turn instruction tuning strategy.

TowerAlign vs. *w/o TowerAlign.* Removing TowerAlign consistently degrades multimodal performance and exacerbates expert selection collapse on visual tokens, confirming that modality-wise norm alignment is a necessary condition for stable MoE training in our setting.

Table 2: Ablation results on MoE-dVLMs. We report the official score for each benchmark. TokRMS applies token-wise RMSNorm at the tower interface; **w/o TowerAlign** uses no post-projector alignment; **LB loss** increases the load-balancing coefficient to 10× (0.1); **Gumbel** injects Gumbel-Softmax noise ($\tau = 2.0$); **Per-turn SFT** trains each turn independently instead of packing a full multi-turn dialogue.

Benchmark	TowerAlign	TokRMS	w/o TowerAlign	Gumbel	LB loss	Per-turn SFT
ChartQA	59.0	19.4	4.7	8.1	9.1	43.8
DocVQA (val)	67.7	26.7	13.8	6.8	9.3	55.3
InfoVQA (val)	25.8	13.4	6.8	5.2	8.1	19.7
RealWorldQA	58.9	36.7	30.6	16.7	28.9	55.6
MMStar	46.7	38.9	39.6	23.1	28.5	45.5
AI2D	70.5	48.2	58.7	28.9	34.9	67.7
MMBench (EN)	73.3	58.0	61.4	48.2	43.9	72.1
MMMU (val)	40.3	31.2	36.7	25.4	31.5	37.3
MMMU-Pro (standard)	25.2	17.5	23.3	15.0	16.9	23.6
MMMU-Pro (vision)	15.1	16.1	14.2	11.3	10.1	14.3

Global Affine Compression vs. *RMSNorm.* We compare our global affine compression to a per-token RMSNorm baseline applied at the same interface. While RMSNorm equalizes per-token magnitudes, it also erases relative norm structure among visual tokens, which can be informative for downstream visual understanding [16]. In practice, GAC yields stronger and more stable performance, supporting our design choice of structure-preserving global rescaling.

TowerAlign vs. *router-side regularization.* We further compare TowerAlign against common routing-side interventions, including stronger load-balancing losses and Gumbel-Softmax perturbations. As shown in Figure 2, these baselines fail to resolve visual-token collapse and can even worsen expert utilization entropy while harming downstream quality, suggesting that the bottleneck lies upstream in representation geometry rather than in router optimization.

Multi-turn supervision: per-turn vs. *joint multi-turn.* Finally, we study instruction tuning with multi-turn conversations: (a) training one turn at a time (treating each turn as an independent sample) versus (b) training the full multi-turn dialogue as a single sequence following LLaDA-V. Note that under bidirectional attention, joint multi-turn training inevitably leaks future-turn context when predicting earlier answers, introducing a train–test mismatch. Surprisingly, we find this strategy yields better overall performance in our setting, likely because the model can exploit cross-turn consistency signals during denoising and learn stronger dialogue-level representations. It is also more efficient in practice: packing multiple turns into a single sequence amortizes the vision encoding/prompt overhead and improves token utilization, reducing the number of forward passes needed per conversation.

7 Conclusion

We investigate why sparse Mixture-of-Experts can degrade when applied to dVLMs, and identify a key failure mode: **expert selection collapse on visual tokens.** Our analysis shows that this collapse is not primarily a router-optimization issue; instead, a severe modality-wise norm discrepancy at the tower interface induces attention drowning and representational inertia in Pre-Norm Transformers, making visual tokens geometrically indistinguishable and leading to degenerate top-k expert assignments.

To address this, we propose **TowerAlign**, a lightweight post-projector norm alignment that rescales visual embeddings to match text norm statistics while preserving relative structure among visual tokens. Across a broad evaluation suite, TowerAlign consistently improves multimodal performance for a 7B-A1B MoE-dVLM and is competitive with dense diffusion baselines under matched data budgets, with negligible additional computation. Extensive ablations further validate our design choices and show that routing-side regularization alone is insufficient, highlighting the importance of upstream geometric alignment. We hope these findings provide a practical recipe and a clearer diagnostic lens for building scalable MoE-dVLMs.

Acknowledgements

We gratefully acknowledge the support from the Distinguished Young Scholars Project funded by the Natural Science Foundation of Guangdong Province (No. 2025B1515020060), and the Basic and Applied Basic Research Program of the Guangzhou Science and Technology Plan (No. 2025A04J7141).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bie, T., Cao, M., Chen, K., Du, L., Gong, M., Gong, Z., Gu, Y., Hu, J., Huang, Z., Lan, Z., et al.: Llada2. 0: Scaling up diffusion language models to 100b. arXiv preprint arXiv:2512.15745 (2025)
4. 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)
5. 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)
6. 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)
7. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
8. Li, B., Zhang, Y., Wang, X., Liang, W., Schmidt, L., Yeung-Levy, S.: Closing the modality gap for mixed modality search (2025), https://arxiv.org/abs/2507.19054
9. Li, B., Xue, X., Yang, S., Shi, Y., Chen, X., Guan, Y., Zhang, Y., Zhang, W.: The unseen bias: How norm discrepancy in pre-norm mllms leads to visual information loss. arXiv preprint arXiv:2512.08374 (2025)
10. Li, B., Xue, X., Yang, S., Shi, Y., Chen, X., Guan, Y., Zhang, Y., Zhang, W.: The unseen bias: How norm discrepancy in pre-norm mllms leads to visual information loss (2025), https://arxiv.org/abs/2512.08374
11. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
12. Li, S., Kallidromitis, K., Bansal, H., Gokul, A., Kato, Y., Kozuka, K., Kuen, J., Lin, Z., Chang, K.W., Grover, A.: Lavida: A large diffusion language model for multimodal understanding. arXiv preprint arXiv:2505.16839 (2025)
13. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge (2024)
14. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
15. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
16. Luo, J., Fan, W.C., Wang, L., He, X., Rahman, T., Abolmaesumi, P., Sigal, L.: To sink or not to sink: Visual information pathways in large vision-language models. arXiv preprint arXiv:2510.08510 (2025)
17. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
18. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
19. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
20. Ni, J., team: Openmoe 2: Sparse diffusion language models. https://github.com/JinjieNi/OpenMoE2 (2025)
21. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models. arXiv preprint arXiv:2502.09992 (2025)
22. Qi, J., Liu, J., Tang, H., Zhu, Z.: Beyond semantics: Rediscovering spatial awareness in vision-language models. arXiv preprint arXiv:2503.17349 (2025)
23. Sun, K., Yuan, X., Liu, H., Zhao, C., Zhang, C., Dredze, M., Bai, F.: Reading, not thinking: Understanding and bridging the modality gap when text becomes pixels in multimodal llms (2026), https://arxiv.org/abs/2603.09095

24. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
25. 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)
26. x.ai: Grok-1.5 vision preview (2024), `https://x.ai/news/grok-1.5v/`, [Online]. Available: `https://x.ai/news/grok-1.5v/`
27. Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., Kong, L.: Dream 7b: Diffusion large language models. arXiv preprint arXiv:2508.15487 (2025)
28. You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.R., Li, C.: Llada-v: Large language diffusion models with visual instruction tuning. arXiv preprint arXiv:2505.16933 (2025)
29. Yu, X., Xin, Y., Zhang, W., Liu, C., Zhao, H., Hu, X., Yu, X., Qiao, Z., Tang, H., Yang, X., et al.: Modality gap-driven subspace alignment training paradigm for multimodal large language models. arXiv preprint arXiv:2602.07026 (2026)
30. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
31. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15134–15186 (2025)
32. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 881–916 (2025)
33. Zhu, F., You, Z., Xing, Y., Huang, Z., Liu, L., Zhuang, Y., Lu, G., Wang, K., Wang, X., Wei, L., et al.: Llada-moe: A sparse moe diffusion language model. arXiv preprint arXiv:2509.24389 (2025)