LookME: Lookup-Based Multimodal Embeddings for Layer Injection in Vision-Language Models

Zeyu Xu*, Xingzhong Hou*, Pengkai Guo, Siling Lin, Xiao Xu, Menghua Zhai, Haoyu Chen, Yunke Zhang†, Fei Huang†

Honor Device Co., Ltd
zhangyunke@honor.com, huangbo6@honor.com

Abstract

Vision-Language Models (VLMs) have achieved strong progress in multimodal understanding. However, scaling dense or sparse Mixture-of-Experts (MoE) models to improve performance limits deployment in resource-constrained environments due to the trade-off between high memory usage from full loading and increased latency from on-demand loading. Recently, the Per-Layer Embedding (PLE) architecture addresses this by scaling models with large external embedding tables stored in ROM and performing lightweight lookup to retrieve relevant embeddings to enhance token representations. Nevertheless, existing PLE-style methods are primarily designed for text embeddings due to the convenience of ID-based retrieval, limiting their effectiveness in VLMs where multimodal embeddings contain richer information for visual tasks. In this paper, we propose LookME, the first framework that enables lookup-based enhancement for multimodal embeddings in VLMs while supporting partitioned storage and on-demand loading. To efficiently lookup arbitrary continuous multimodal embeddings from large-scale embedding tables, we propose a hierarchical two-level lookup method employing a coarse-to-fine strategy that performs lookups from the scene-level to the intra-scene primitive-level. Furthermore, we integrate the lookup method with a sparse injection strategy, which adaptively prioritizes critical embeddings over voluminous multimodal embeddings within layers, and facilitates embedding table reuse across neighboring layers, improving the trade-off among efficiency, model size, and performance. Experiments on multiple visual benchmarks show that LookME outperforms text-only PLE-style methods, validating the effectiveness of lookup-based multimodal embedding enhancement.

Introduction

Vision-Language Models (VLMs) have become a dominant paradigm for multimodal intelligence, achieving remarkable performance across diverse visual tasks (Liu et al. 2023; Li et al. 2023; Google DeepMind 2023; Chen et al. 2023). Following scaling laws (Alayrac et al. 2022; Kaplan et al. 2020), recent VLMs have expanded to 72B parameters or larger (Bai et al. 2023; Chen et al. 2024d; Team et al. 2025). However, dense scaling incurs prohibitive FLOPs, bandwidth, and energy costs (Hoffmann et al. 2022). Mixture-of-Experts (MoE) methods activate only a fraction of parameters via token routing (Shazeer et al. 2017; Fedus, Zoph, and Shazeer 2022), enabling larger scales with low computational cost. However, MoE typically requires sufficient GPU memory to store the entire model. Otherwise, frequent loading of large, non-contiguous expert weights introduces significant latency, hindering deployment on resource-constrained edge devices (Wu et al. 2024).

Recently, Per-Layer Embedding (PLE) has been proposed as an alternative paradigm for sparse parameter activation (Google DeepMind 2025). Instead of directly increasing the number of model parameters, PLE augments the model with large-scale token-level external knowledge embedding tables. During inference, PLE performs lightweight lookups to integrate relevant embeddings for information enhancement. By storing knowledge tables on offline media (e.g., ROM) (Ding et al. 2026) rather than GPU memory, PLE expands model capacity while remaining ideal for resource-constrained edge devices. However, existing PLE-based methods are primarily designed for large language models (LLMs) (Cheng et al. 2026; Sadhukhan et al. 2026; Liu et al. 2026), as discrete text tokens naturally facilitate efficient lookup operations. Although some multimodal models adopt PLE (Google DeepMind 2025), the enhancement is typically restricted to text tokens. In contrast, VLMs are usually dominated by visual tokens (Dosovitskiy et al. 2021), with text tokens comprising only a minor fraction. Consequently, text-only PLE yields marginal gains for multimodal understanding, as the primary multimodal information pathway remains largely unenhanced.

In this paper, we propose LookME, the first framework enabling lookup-based enhancement for multimodal embeddings in VLMs. As illustrated in Figure 1, LookME supports partitioned storage and on-demand loading, extending lookup mechanisms beyond text to the dominant multimodal stream. Since multimodal embeddings lack inherent discrete IDs, we introduce a hierarchical two-level lookup method, which first coarsely routes embeddings to individual scenes, then performs fine-grained retrieval based on the relevance to scene-primitives. This design ensures balanced lookups while preventing collapse for a large-scale embedding table, naturally extending to all multimodal embeddings like multi-

*These authors contributed equally.
†Corresponding author.

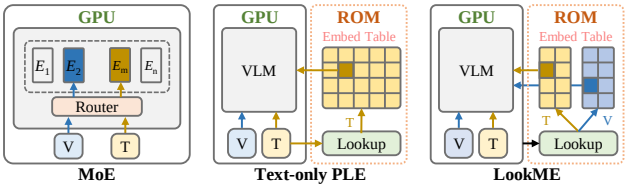

Figure 1: Comparison of architectures between different sparse models.

scale image embeddings and cross-modal embeddings. Furthermore, to address redundancy and parameter overhead, we propose a sparse injection strategy that adaptively selects critical embeddings for lookup within layers and dynamically reuses tables across layers, enhancing both efficiency and performance. Experiments on multiple benchmarks demonstrate that LookME consistently outperforms baselines and text-only PLE-style methods. Our main contributions can be summarized as:

- We propose LookME, a framework that extends PLE-style methods from text to multimodal embeddings, enabling offline embedding construction, lookup, and enhancement for dominant multimodal information in VLMs.
- We design a multimodal embedding lookup method that employs ID mapping for discrete text embeddings and a hierarchical two-level lookup method for continuous image and cross-modal embeddings.
- We design a sparse injection strategy that adaptively injects looked-up embeddings within and across layers, striking a balance between efficiency, model size, and performance.
- Experiments on multiple visual task benchmarks demonstrate that LookME outperforms baselines as well as text-only PLE-style methods.

Related Works

Vision-Language Models

In recent years, VLMs have attracted significant attention due to their strong capabilities in multimodal understanding (Wang et al. 2023; Zhang et al. 2022). Empirical evidence consistently supports the scaling law, where larger VLMs tend to exhibit stronger visual understanding abilities. Consequently, the parameter scale of VLMs has steadily increased. Early models such as LLaVA and InstructBLIP typically operated at scales ranging from 7B to 13B parameters (Liu et al. 2023; Dai et al. 2023). Subsequent work adopted progressively larger backbone architectures; for example, Qwen-VL, InternVL, and DeepSeek-VL expanded model sizes to 72B parameters or beyond (Bai et al. 2023; Chen et al. 2024d; Lu et al. 2024). Despite the notable performance gains brought by scaling, this trend also incurs substantial computational and memory costs (Zhang et al. 2025b), creating significant challenges for deployment in resource-constrained environments such as edge devices. This growing tension between model capacity and inference efficiency motivates the exploration of parameter-efficient approaches that can enhance multimodal capability without proportionally increasing the number of active parameters (Zhang et al. 2025a).

Sparse Parameter Scaling

Sparse parameter scaling aims to expand model capacity while activating only a subset of parameters, reducing computational overhead. As a primary paradigm, MoE dynamically routes tokens to expert subsets, achieving strong performance in LLMs and VLMs (Dai et al. 2024). However, MoE typically requires all parameters to be loaded on-device to avoid switching latency, leading to substantial memory overhead that limits its use on resource-constrained devices. To address this challenge, Gemma-3 (Google DeepMind 2025) proposes a parameter-efficient lookup method termed PLE, which leverages ROM-stored embedding tables indexed by vocabulary IDs. This approach strikes a balance between memory constraints and efficiency while maintaining high quality. MeKi (Ding et al. 2026) also equips LLM layer with token-based embeddings and applies a re-parameterization strategy to offload these parameters into a compact static lookup table stored in ROM, supporting loading with low latency. Engram (Cheng et al. 2026) and LongCAT (Liu et al. 2026) further extend PLE from single text token to N-gram using specialized hashing methods, They integrate the retrieved embeddings into the backbone model through contextualized gating, achieving performance competitive with MoE. STEM (Sadhukhan et al. 2026) replaces FFN up-projections with token-indexed lookups, enabling efficient sparse scaling with improved stability and reduced overhead.

However, existing PLE-style methods are primarily designed for LLMs and text tokens, limiting their direct applicability to VLMs. In this paper, we propose LookME, which extends the lookup mechanism to multimodal embeddings, thereby enhancing the dominant information within VLMs.

Methods

Overview

Text-only PLE methods enhance the hidden states $H_{\text{in}} \in \mathbb{R}^{L \times d_{\text{h}}}$ at each predefined layer as

$$\begin{aligned} H_{\text{out}} &= \text{DecoderLayer}(H_{\text{in}}) + \mathcal{G}(\mathcal{R}_{\text{text}}), \\ \mathcal{R}_{\text{text}} &= E[u], \end{aligned} \tag{1}$$

where L denotes the sequence length, d_{h} denotes the hidden dimension of the decoder layers, $\mathcal{R}_{\text{text}}$ the retrieved external embeddings using the input text token IDs u from an external embedding table $E \in \mathbb{R}^{|V| \times d_{\text{e}}}$ with $|V|$ representing the vocabulary size, d_{e} the embedding dimension, and $\mathcal{G}$ the injection function. The process leaves the dominant multimodal tokens in VLMs untouched and operates on all positions indiscriminately, regardless of their heterogeneous signal enhancement requirements.

To address these limitations, we propose LookME, as illustrated in Figure 2. LookME layers $\mathcal{L}_{\text{aug}}$ are integrated into intermediate decoder layers, supporting *Text*, *Image*, and *Cross-Modal* as three complementary pathways for lookup. By employing sparse injection, $\mathcal{L}_{\text{aug}}$ selectively augments the decoder with retrieved multimodal external embedding

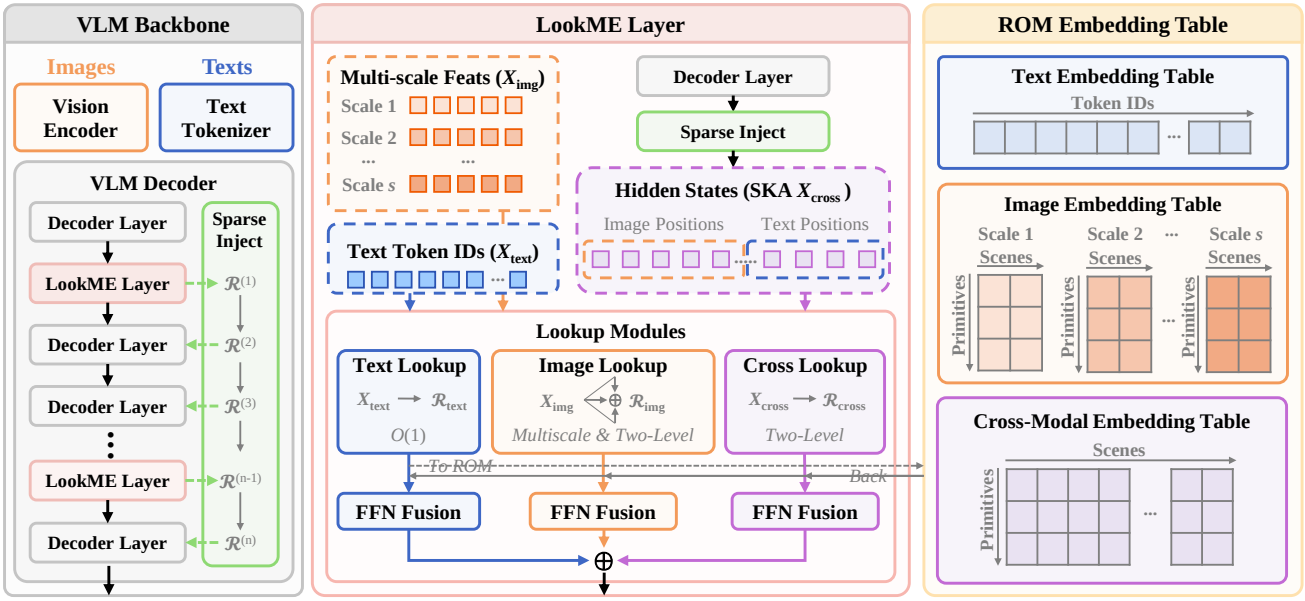

Figure 2: The overall workflow of LookME. External embedding tables are separately stored in ROM from the main backbone. LookME layers are inserted into the decoder, where consecutive layers communicate and share embedding tables via sparse intra-layer propagation. Each LookME layer includes an SKA gate for sparse inter-layer activation and three parallel lookup pathways to retrieve embeddings for enhancement.

across tokens and layers,

$$H_{\text{out}} = \mathcal{L}_{\text{aug}}(H_{\text{in}}) = \text{DecoderLayer}(H_{\text{in}}) + \mathcal{S}(\mathcal{R}_{\text{text}} + \mathcal{R}_{\text{img}} + \mathcal{R}_{\text{cross}}), \quad (2)$$

where $\mathcal{R}_*$ denotes the per-modality pathway retrieval (in Section Multimodal Embedding Lookup) and $\mathcal{S}$ denotes the sparse injection (in Section Sparse Injection).

Multimodal Embedding Lookup

For a given LookME layer $\mathcal{L}_{\text{aug}}^{(l)}$, the multimodal embedding lookup module retrieves external embeddings through three complementary pathways to enhance H_{in}. Each pathway $\mathcal{R}_*$ contains an individual input X_* and a retrieval function $\mathcal{T}_*$ that retrieves embeddings from its external embedding table E_*,

$$\begin{aligned} \mathcal{R}_{\text{text}}^{(l)} &= \mathcal{T}_{\text{text}}(X_{\text{text}};\ E_{\text{text}}^{(l)}), \\ \mathcal{R}_{\text{img}}^{(l)} &= \mathcal{T}_{\text{img}}(X_{\text{img}};\ E_{\text{img}}^{(l)}), \\ \mathcal{R}_{\text{cross}}^{(l)} &= \mathcal{T}_{\text{cross}}(X_{\text{cross}}^{(l)};\ E_{\text{cross}}^{(l)}). \end{aligned} \quad (3)$$

The external embedding tables E_* for each pathway are initialized using principal component analysis (PCA) derived from the pretrained backbone embedding matrix, providing improved initialization while preserving a shared latent subspace across modalities. The input X_* and retrieval function $\mathcal{T}_*$ vary across pathways. Specifically, the text pathway uses discrete token IDs with direct lookup and is only applied to text positions; the image pathway uses continuous multiscale visual encoder embeddings with a hierarchical two-level lookup method and is only applied to image positions; the cross-modal pathway uses continuous decoder hidden states with a hierarchical two-level lookup method and is applied to all positions. Details of each pathway are described below.

Text Embedding Index-Based Lookup The text pathway is applied to the text positions in the original combined image-query input. It directly retrieves embeddings from the text external embedding table using the corresponding text token IDs $E_{\text{text}}[u]$, following the text-only PLE methods. This $O(1)$ lookup injects a context-independent lexical prior into the text positions.

Image Embedding Hierarchical Two-Level Lookup The image pathway is applied to the image positions in the original combined image-query input. The image embeddings $v \in \mathbb{R}^{L_{\text{v}} \times d_{\text{v}}}$ generated by the visual encoder are used for retrieval, where L_{v} and d_{v} denote the number and dimensionality of image embeddings, respectively. These embeddings are directly available without additional computation and provide representative and stable retrieval signals. Since continuous image embeddings lack discrete vocabulary IDs for direct indexing, the retrieval process is fundamentally formulated either as a similarity-based matching (via a projection $p \in \mathbb{R}^{d_{\text{v}} \times d_{\text{e}}}$ against the embedding table $E_{\text{img}} \in \mathbb{R}^{|V| \times d_{\text{e}}}$) or as a classification task (via a weight matrix $w \in \mathbb{R}^{d_{\text{v}} \times |V|}$). As the visual vocabulary expands to a text-equivalent scale, the computational overhead becomes exorbitant, and the retrieval process is prone to severe routing imbalance, resulting in massive parameter under-utilization. To overcome this limitation, we design a hierarchical two-level retrieval method. Image embeddings are initially routed to a coarse-grained *scene*, followed by a dynamic soft-combination of specific

scene primitives to enrich details. Details are presented in the following sections.

- **Two-level embedding table.** A flat $|V|$-way routing distribution is statistically sparse and prone to imbalance. For hierarchical two-level retrieval, we reorganize the image external embedding table into a two-dimensional structure

$$E_{\text{img}} \in \mathbb{R}^{N_{\text{s}} \times N_{\text{p}} \times d_{\text{e}}}, \tag{4}$$

with N_{s}, N_{p} denoting the numbers of *scenes* and *scene-primitives*, respectively, and $N_{\text{s}} \times N_{\text{p}}$ being comparable to $|V|$. The scene axis governs where to look (a dense, discrete N_{s}-way decision); the scene-primitive axis governs what to read (a sparse, differentiable combination). This factorization provides statistical stability at the scene level and fine-grained expressivity at the scene-primitive level, without increasing the number of parameters compared with a flat table of the same total size.

- **Level-1: scene retrieval.** Level-1 performs coarse-grained scene-level retrieval by categorizing image embeddings into different scenes according to their global semantics, thereby facilitating subsequent fine-grained retrieval within the selected scene. We use scene-centered embeddings $e_{\text{img}} \in \mathbb{R}^{N_{\text{s}} \times d_{\text{e}}}$ to represent scene features, which are dynamically updated during training

$$e_{\text{img}} = \frac{1}{N_{\text{p}}} \sum_{p=1}^{N_{\text{p}}} E_{\text{img}}[:, p, :]. \tag{5}$$

The image embedding is then routed to the corresponding scene based on cosine similarity:

$$S = \arg\max \left(\cos(v_{\text{e}}, e_{\text{img}}) + b \right), \tag{6}$$

where v_{e} denotes the image embeddings projected into the external embedding space via $W_{\text{v}\to\text{e}} \in \mathbb{R}^{d_{\text{e}} \times d_{\text{v}}}$, and $b \in \mathbb{R}^{N_{\text{s}}}$ is a per-scene router balancing bias adaptively updated from the discrepancy between empirical routing frequencies and predicted probabilities (DeepSeek-AI 2024).

- **Level-2: scene-primitive retrieval.** Level-1 routes image embeddings into coarse scenes, while Level-2 supplements fine-grained details by generating soft combination weights over the scene-primitives within each scene. Through the dynamic composition of scene-primitives, Level-2 can synthesize an external embedding that is most relevant to the input under the current scene context. Unlike Level-1, Level-2 excludes scene-primitives from the retrieval process to mitigate the parameter loading challenges similar to MoE, which arise when tokens target diverse scenes whose primitives may, in the worst case, encompass the entire external table, thereby necessitating a trade-off between on-demand loading latency and one-time preloading memory overhead. Thus, we leverage a shared two-layer MLP to estimate the relevance scores of scene primitives s_p based on the projected image embeddings v_e and scene-centered embeddings e_{img} for a selected scene S, and retrieve the top-$\tilde{T}$ primitives $e_{S,p}$

$$s_p = \text{MLP}(v_{\text{e}}, e_{\text{img}}[S, :]), \quad e_{S,p} = \text{argTop}\tilde{T}\{s_p\}. \tag{7}$$

These primitives are then integrated into external image embeddings based on relevance scores and similarity

$$\mathcal{R}_{\text{img}} = \sum_{p \in \mathcal{I}_{\tilde{T}}} \phi \left(\cos(v_{\text{e}}, e_{S,p}) + \tau(s_p) \right) \cdot e_{S,p}, \tag{8}$$

where ϕ and τ are softmax and tanh functions, respectively.

- **Multi-scale image retrieval.** To capture visual features at different granularities, we perform the hierarchical two-level retrieval independently on image embeddings extracted from multiple intermediate layers of the visual encoder, each equipped with its own table and coarse ranker. The retrieved embeddings from different scales are fused at each image position into a single representation through a token-wise softmax router conditioned on the input sequence H_{in}, enabling to determine the most suitable combination of granularities for each token,

$$\begin{gathered} \mathcal{R}_{\text{img}} = \phi \left((H_{\text{img}} W_{\text{h}\to\text{e}} \cdot \tilde{\mathcal{R}}_{\text{img}}^{\top}) / \sqrt{d_{\text{e}}} \right) \tilde{\mathcal{R}}_{\text{img}}, \\ \tilde{\mathcal{R}}_{\text{img}} = [\mathcal{R}_{\text{img},1}, \ldots, \mathcal{R}_{\text{img},s}], \end{gathered} \tag{9}$$

where $W_{\text{h}\to\text{e}} \in \mathbb{R}^{d_{\text{e}} \times d_{\text{h}}}$ is a projection matrix, H_{img} denotes the image-position hidden states of H_{in} after the decoder layer, $\mathcal{R}_{\text{img},s}$ represents the retrieved embeddings using the img embedding at scale s.

Cross-Modal Embedding Hierarchical Two-Level Lookup The cross-modal pathway is applied to all position. The retrieval input X_{c} is obtained from H_{in} after the decoder in each LookME layer, as they directly provide the cross-modal information. Similar to the image pathway, the cross-modal pathway also adopts the hierarchical two-level lookup method to retrieve continuous cross-modal external embeddings, where each component in the method is independent of the image pathway. The cross-modal pathway provides a unified mechanism for retrieving external embeddings at arbitrary positions. Unlike single-modality pathways that operate on modality-specific token subsets, it enables flexible embedding integration across text, image, audio, or any combination of modalities.

Sparse Injection

For a given LookME layer $\mathcal{L}_{\text{aug}}^{(l)}$, the multimodal embedding lookup module retrieves three embeddings $\{\mathcal{R}_{\text{text}}^{(l)}, \mathcal{R}_{\text{img}}^{(l)}, \mathcal{R}_{\text{cross}}^{(l)}\}$, which are subsequently integrated through the injection function $\mathcal{S}$, as defined in Eq. 2. However, both retrieval and injection introduce additional computational overhead, particularly for the image and cross-modal pathways with long sequences. Moreover, converting every decoder layer into a LookME layer significantly increases the model size due to the dedicated multimodal embedding tables. Thus, we propose a sparse injection method to address these issues separately. The intra-layer sparse activation module generates a per-token mask to filter out low-signal tokens, retaining only informative tokens for retrieval and injection (in Section Intra-Layer Sparse Activation). The inter-layer sparse activation module converts only selected decoder layers into LookME layers and introduces a propagation network that carries retrieved embeddings across non-LookME

layers, enabling shallow-layer knowledge to remain accessible at deeper layers without repeated retrieval (in Section Inter-Layer Sparse Activation). Details are described below.

Intra-Layer Sparse Activation The intra-layer sparse activation is controled by sparse knowledge activation (SKA), which is a per-token, per-pathway binary gate to decides whether to invoke the embedding pathway at position i. SKA gate is computed as:

$$\begin{aligned} g_i &= \sigma(W_{\mathrm{g}}\Delta_i), \\ \Delta_i &= \mathrm{Attn}(H_{\mathrm{in}})_i, \end{aligned} \tag{10}$$

where $W_{\mathrm{g}} \in \mathbb{R}^{d_{\mathrm{h}} \times 1}$ denotes the gate projection, σ represents the Gumbel-Sigmoid function, which remains soft during training for differentiability and exploration while reducing to hard during inference. Δ_i denotes the attention contribution of the host LookME layer, specifically the additive update produced by self-attention at position i before the residual connection in decoder layer. Δ_i is obtained at no additional computational cost, and tokens with large $|\Delta_i|$ are substantially updated by self-attention and are therefore more likely to benefit from external embedding augmentation, whereas tokens with near-zero values remain largely unchanged and receive limited benefit from additional embeddings. By conditioning on the shared attention contribution while applying distinct gate projections for multimodal pathways, the SKA gate is able to adaptively learn suitable pathway-specific activation patterns.

Each pathway in the multimodal embedding lookup module is executed only on the activated positions:

$$\mathcal{S}(\mathcal{R}_*) = g \odot f\big(\mathcal{R}_*[g]\big), \tag{11}$$

where f denotes a SwiGLU-style FFN fusion that projects the retrieved embeddings from dimension d_{e} to the backbone dimension d_{h}, and $\mathcal{S}(\mathcal{R}_*)$ is further fused with the backbone sequentially as Eq. (2).

Inter-Layer Sparse Activation Inter-layer sparse activation selectively transforms only a subset of decoder layers into LookME layers, thereby avoiding the substantial parameter overhead of a full conversion while still allowing non-LookME layers to benefit from the enhancement. Considering that adjacent layers tend to retrieve similar scene primitives, we convert only a small subset of layers into LookME layers, while allowing each intermediate non-LookME layer to inherit the sparse injection output from the preceding LookME layer through a lightweight bottleneck propagation, instead of re-running the full retrieval process.

Assuming k is a non-LookME layer and l is the nearest LookME layer above k, the sparse activation for the non-LookME layer k is computed as

$$\begin{aligned} \mathcal{S}^{(k)}({\mathcal{R}_*}^{(k)}) &= g^{(l)} \odot f^k\big({\mathcal{R}_*}^{(k)}[g^{(l)}]\big) \\ {\mathcal{R}_*}^{(k)} &= {\mathcal{R}_*}^{(k-1)} W_{\mathrm{p}}^k + {\mathcal{R}_*}^{(k-1)} \end{aligned} \tag{12}$$

where $W_{\mathrm{p}}^k \in \mathbb{R}^{d_{\mathrm{e}} \times d_{\mathrm{e}}}$ denotes the propagation projection in the embedding space, which is more lightweight than the backbone space since $d_{\mathrm{e}} \ll d_{\mathrm{h}}$. Through inter-layer sparse activation, enhancements to the backbone network become significantly more flexible and adaptable.

Experiments

In this section, we evaluate LookME against state-of-the-art methods, and further provide visual analyses of the lookup process together with comprehensive ablation studies. Additional experimental results are included in the supplementary material.

Experimental Setup

Training Details Our training corpus is constructed entirely from open-source image–text datasets, including but not limited to (Gu et al. 2024; Wiedmann et al. 2025; Tong et al. 2024; Yuan et al. 2025; Chen et al. 2024c,a; Deitke et al. 2024). We curate and sample the data according to representative vision tasks to maintain a balanced distribution across diverse capabilities. The resulting dataset contains 60 million samples, comprising approximately 57 billion visual tokens and 14 billion text tokens. To ensure fair comparison, all models in this paper are pre-trained on the same unified dataset.

We adopt Qwen2.5VL-3B (Bai et al. 2025) as the backbone, as it provides strong multimodal performance while remaining suitable for deployment on edge devices. Based on this backbone, we perform full-parameter fine-tuning to train LookME and all corresponding baselines. Detailed training configurations and hyperparameters are provided in the supplementary material.

Evaluation Details We compare LookME against three categories of baselines. First, the original Qwen2.5VL-3B serves as a general baseline. Second, we include Qwen2.5VL-3B further pre-trained on our unified dataset to isolate the gains brought by the training data itself. Third, we compare with state-of-the-art text-only memory-/knowledge-enhanced methods, including Engram (Cheng et al. 2026) and MeKi (Ding et al. 2026), using the same hyperparameter settings as LookME to validate the effectiveness of our multimodal embedding lookup framework.

To comprehensively evaluate performance, we conduct experiments on 11 established benchmarks covering diverse visual tasks using VLMEvalKit (Duan et al. 2024), including *STEM*: MMMU (Yue et al. 2024) and MathVision (Wang et al. 2024a); *General VQA*: MMBench-EN (Liu et al. 2024a) and MMStar (Chen et al. 2024b); *Hallucination*: HallusionBench (Guan et al. 2024); *Document Understanding*: InfoVQA (Mathew et al. 2021) and OCRBench (Liu et al. 2024b); *Visual Perception*: HRBench8K (Wang et al. 2024b), CV-Bench-3D and CCBench (Liu et al. 2024a); *Multi-Image*: BLINK (Fu et al. 2024).

Main Results

Comparison with SOTA Methods We compare LookME with four representative methods: the vanilla Qwen2.5-VL and Qwen2.5-VL$_{\mathrm{CPT}}$ baselines, as well as two recent state-of-the-art memory-/knowledge-enhanced methods, Engram and MeKi. Table 1 summarizes the results on 11 benchmarks spanning six task categories. The results demonstrate that integrating retrieved information directly into the backbone's multimodal representation stream leads to consis-

Table 1: Comparison with SOTA methods. The best results are in **bold** and the second best are underlined. All models are evaluated under the same zero-shot settings.

Task	Benchmark	Models				
		Qwen25-VL	**Qwen25-VL$_{CPT}$**	**Engram**	**MeKi**	**LookME**
LLM Params (Backbone / External)		3.09B / -	3.09B / -	3.09B / 2.35B	3.09B / 0.26B	3.09B / 0.82B
STEM	MMMU	46.9	47.1	41.8	46.3	**48.7**
	MathVision	17.9	18.1	17.4	18.2	**20.4**
General VQA	MMBench-EN	78.3	78.7	73.7	77.1	**79.8**
	MMStar	55.1	55.7	54.2	54.5	**57.3**
Hallucination	HallusionBench	65.2	65.8	68.0	69.2	**70.1**
Document Understanding	InfoVQA	75.5	76.1	69.1	74.4	**77.2**
	OCRBench	82.9	83.7	82.5	82.7	**85.1**
Visual Perception	HRBench8K	63.3	63.4	60.9	62.9	**66.5**
	CV-Bench-3D	74.8	75.3	71.3	73.6	**79.2**
	CCBench	62.6	63.7	61.6	63.7	**66.1**
Multi-Image	BLINK	48.6	47.6	47.8	**49.0**	48.7

tent improvements over both standard baselines and prior memory-/knowledge-enhanced approaches.

As shown in Table 1, LookME consistently achieves the best performance across all six task categories, with the performance gap becoming more pronounced on tasks that require stronger fine-grained multimodal alignment. On standard single-image benchmarks, the improvement over the strongest baseline is modest (e.g., **+1.6** on MMMU). In contrast, on more challenging visual perception benchmarks, the gain increases to **+2.8**–**+3.9** points (e.g., CV-Bench-3D: **79.2** vs. 75.3). Notably, previous memory-/knowledge-enhanced methods such as Engram and MeKi underperform the Qwen2.5-VL backbone on most benchmarks.

We attribute these results to the fundamentally multimodal design of LookME. Unlike prior methods that operate primarily in the text domain and largely neglect visual representations, LookME explicitly models cross-modal interactions and incorporates retrieved information within the visual-language representation space itself. This advantage becomes especially important in scenarios requiring evidence aggregation and alignment across multiple images, where prior approaches struggle and LookME demonstrates its largest gains.

Ablation Experiment

Effectiveness of Multimodal Embedding Lookup A key design of LookME is that the embedding lookup is driven by *multiple modalities* rather than text alone. In the default configuration, we construct three query pathways—text, image, and hidden states—whose retrieved representations are jointly integrated for subsequent augmentation. To evaluate the contribution of each pathway, we perform an ablation study over different modality combinations, with results reported in Table 2.

Specifically, we compare three progressively richer variants: *Text-only*, *Text+Img*, and the full *Text+Img+Hidden* configuration. As shown in Table 2, performance improves consistently as additional modalities are introduced. The *Text-only* variant achieves the weakest results, while incorporating the image pathway yields clear gains across all four benchmarks (e.g., +1.9 on HRBench and +0.6 on MMMU). The complete *Text+Img+Hidden* model further delivers the best overall performance.

Table 2: Ablation study on multimodal embedding lookup pathways. ✓ / × denote whether the pathway is enabled or removed.

Modalities			Benchmarks			
T	**I**	**C**	**MMMU**	**MMB.**	**OCRB.**	**HRB.**
✓	×	×	47.1	78.2	83.7	63.9
✓	✓	×	47.7	79.3	84.6	65.8
✓	✓	✓	**48.7**	**79.8**	**85.1**	**66.5**

These results indicate that each query pathway contributes complementary information to the retrieval process. More importantly, they support the core motivation behind LookME: unlike prior memory-/knowledge-enhanced approaches that operate primarily in the text-only regime, LookME explicitly constructs multimodal query pathways that remain aligned with the backbone's joint visual-textual representation space. This multimodal alignment is a key factor underlying LookME's performance advantage.

Effectiveness of Sparse Injection LookME introduces sparse injection at two complementary granularities: *within-layer* sparsity through Sparse Kernel Activation (SKA), and *across-layer* propagation through NetProp. To assess their individual contributions, we conduct separate ablations for each component, and discuss the corresponding results in the

Table 3: Ablation study on across-layer sparse injection.

NetProp Range	MMMU	MMB.	OCRB.	HRB.
None	47.0	79.7	84.9	65.0
Full	47.8	78.8	84.3	**67.6**
Head-Tail	**48.7**	**79.8**	**85.1**	66.5

Table 4: Ablation study on the embedding dimension of retrieved content.

Dim	MMMU	MMB.	OCRB.	HRB.
128D	47.0	78.7	84.1	64.5
256D	47.7	78.4	84.3	65.8
512D	**48.7**	**79.8**	**85.1**	**66.5**

following paragraphs.

Across-layer: NetProp. We compare three variants of connecting the augmented layers: *None* (no NetProp), *Full* (NetProp in every unaugmented layers), and *Head-Tail* (only between the first and the last augmented layers). As shown in Table 3, *Head-Tail* achieves the best performance on three of the four benchmarks, while *None* consistently performs the worst. Although *Full* attains the highest score on HRBench, it leads to noticeable degradation on MMBench and OCRBench. These results suggest that NetProp is most effective when propagating information between layers with relatively aligned feature representations. Over longer propagation distances, feature distributions drift substantially, making reused keys progressively less reliable. The *Head-Tail* design therefore provides the best balance: it preserves a lightweight global communication path between the entrance and exit of the augmented region, while allowing intermediate layers to retrieve updated information based on their own local features.

Hyperparameter Analysis LookME involves two key design choices: the embedding dimension of the retrieved content and the insertion layers within the backbone. By default, we use an embedding dimension of **512** and inject the retrieved features at layers [2/10/18]. In this section, we conduct an ablation study to evaluate the sensitivity of the model to these two factors. The corresponding results are reported in Tables 4 and 5.

Embedding dimension. We evaluate three embedding dimensions—128, 256, and 512—while keeping the injection layers fixed. As shown in Table 4, the 512-dimensional setting consistently delivers the best performance across all four benchmarks. Reducing the dimension to 256 and 128 leads to a steady decline in performance on every benchmark, with the largest drops observed on HRBench (66.5 $\rightarrow$ 64.5) and MMMU (48.7 $\rightarrow$ 47.0). This trend highlights the importance of sufficient embedding capacity: the retrieved representations must encode diverse visual and textual information, including objects, attributes, relations, and OCR cues. Lower-dimensional embeddings limit the discriminability of the memory space, making reliable retrieval under varied visual contexts more difficult. In contrast, a 512-dimensional embedding provides adequate representational capacity while remaining computationally and memory efficient.

Table 5: Ablation study on the layers where retrieved content is injected.

Layers	MMMU	MMB.	OCRB.	HRB.
18/26/34	46.4	77.2	84.4	64.7
2/18/34	46.7	78.7	83.7	65.6
2/10/18	**48.7**	**79.8**	**85.1**	**66.5**

Injection layers. We compare three injection strategies under the same overall budget: the default shallow-focused configuration [2/10/18], the full-range configuration [2/18/34], and the deep-only configuration [18/26/34]. As reported in Table 5, both [18/26/34] and [2/18/34] consistently underperform the default [2/10/18] across all benchmarks. The deep-only strategy exhibits the largest degradation on MMMU and MMBench, while the full-range setting suffers a notable drop on OCRBench. This behavior can be attributed to the functional roles of different backbone layers. Deeper layers in vision-language models primarily encode high-level semantic interactions, where representations are already highly abstract and relatively saturated, leaving limited capacity for externally retrieved content to meaningfully influence downstream behavior. In contrast, earlier layers preserve richer fine-grained visual and textual features that remain sufficiently plastic to absorb and propagate auxiliary information. Consequently, the default [2/10/18] configuration allocates the injection budget to the shallow and intermediate stages, where retrieved content can more effectively shape the subsequent visual-textual reasoning process.

Limitation LookME is continually pre-trained from a foundation model. We believe that training it from scratch on more comprehensive and diverse multimodal data could further enhance knowledge retention and memory capabilities, ultimately leading to improved performance.

Conclusions

In this paper, we proposed LookME, the first method that extends PLE lookup-based enhancement from text to multimodal embeddings in VLMs, addressing the limitation that existing PLE-style methods only enhance text tokens while leaving dominant multimodal information untouched. LookME introduces a hierarchical two-level lookup method that first coarsely routes continuous multimodal embeddings to scenes, then performs fine-grained retrieval via soft combination of scene-primitives, enabling efficient and balanced lookup at text-equivalent vocabulary scales. A sparse injection strategy further adaptively activates lookup based on attention signals and propagates retrieved embeddings across layers through lightweight transitions, achieving a favorable trade-off among efficiency, model size, and performance. Extensive experiments demonstrate that LookME consistently outperforms baselines and text-only PLE methods across diverse visual benchmarks. LookME supports offline table storage and flexible pathway configuration, offering a promising

direction for parameter-efficient multimodal model expansion.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Chen, G. H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; and Wang, B. 2024a. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. arXiv:2402.11684.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024b. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330*.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salzmann, M.; Beyer, L.; et al. 2023. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024c. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2024d. 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 (CVPR)*. Oral.

Cheng, X.; Zeng, W.; Dai, D.; Chen, Q.; Wang, B.; Xie, Z.; Huang, K.; Yu, X.; Hao, Z.; Li, Y.; Zhang, H.; Zhang, H.; Zhao, D.; and Liang, W. 2026. Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models. *arXiv preprint arXiv:2601.07372*.

Dai, D.; Deng, C.; Zhao, C.; Xu, R. X.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; Xie, Z.; Li, Y. K.; Huang, J.; Luo, F.; Ruan, C.; Sui, Z.; and Liang, W. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv preprint arXiv:2401.06066*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; Lu, J.; Anderson, T.; Bransom, E.; Ehsani, K.; Ngo, H.; Chen, Y.; Patel, A.; Yatskar, M.; Callison-Burch, C.; Head, A.; Hendrix, R.; Bastani, F.; VanderBilt, E.; Lambert, N.; Chou, Y.; Chheda, A.; Sparks, J.; Skjonsberg, S.; Schmitz, M.; Sarnat, A.; Bischoff, B.; Walsh, P.; Newell, C.; Wolters, P.; Gupta, T.; Zeng, K.-H.; Borchardt, J.; Groeneveld, D.; Nam, C.; Lebrecht, S.; Wittlif, C.; Schoenick, C.; Michel, O.; Krishna, R.; Weihs, L.; Smith, N. A.; Hajishirzi, H.; Girshick, R.; Farhadi, A.; and Kembhavi, A. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. arXiv:2409.17146.

Ding, N.; Liu, F.; Kim, K.; Hao, L.; Lee, K.-H.; Ko, H.; and Tang, Y. 2026. MeKi: Memory-based Expert Knowledge Injection for Efficient LLM Scaling. *arXiv preprint arXiv:2602.03359*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11198–11201.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research (JMLR)*, 23(120): 1–39.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. BLINK: Multimodal Large Language Models Can See but Not Perceive. *arXiv preprint arXiv:2404.12390*.

Google DeepMind. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Google DeepMind. 2025. Gemma 3n: On-Device Multimodal AI with Per-Layer Embeddings. Technical report, Google DeepMind.

Gu, S.; Zhang, J.; Zhou, S.; Yu, K.; Xing, Z.; Wang, L.; Cao, Z.; Jia, J.; Zhang, Z.; Wang, Y.; Hu, Z.; Zhang, B.-W.; Li, J.; Liang, D.; Zhao, Y.; Ao, Y.; Liu, Y.; Feng, F.; and Liu, G. 2024. Infinity-MM: Scaling Multimodal Performance with Large-Scale and High-Quality Instruction Data. arXiv:2410.18558.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; Manocha, D.; and Zhou, T. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14375–14385.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; Hennigan, T.; Noland, E.; Millican, K.; van den Driessche, G.; Damoc, B.; Guy, A.; Osindero, S.; Simonyan, K.; Elsen, E.; Rae, J. W.; Vinyals, O.; and Sifre, L. 2022. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, H.; Zhang, J.; Wang, C.; Hu, X.; Lyu, L.; Sun, J.; Yang, X.; Wang, B.; Li, F.; Qian, Y.; Si, L.; Sun, Y.; Li, R.; Pei, P.; Xie, Y.; and Cai, X. 2026. Scaling Embeddings Outperforms Scaling Experts in Language Models. *arXiv preprint arXiv:2601.21204*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024a. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Liu, Y.; Li, Z.; Huang, M.; Yang, B.; Yu, W.; Li, C.; Yin, X.-C.; Liu, C.-L.; Jin, L.; and Bai, X. 2024b. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12).

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; Sun, Y.; Deng, C.; Xu, H.; Xie, Z.; and Zhang, C. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525*.

Mathew, M.; Bagal, V.; Tito, R. P.; Karatzas, D.; Valveny, E.; and Jawahar, C. V. 2021. InfographicVQA. Version Number: 2.

Sadhukhan, R.; Cao, S.; Dong, H.; Zhao, C.; Purpura-Pontoniere, A.; Tian, Y.; Liu, Z.; and Chen, B. 2026. STEM: Scaling Transformers with Embedding Modules. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Team, V.; Hong, W.; Yu, W.; Gu, X.; Wang, G.; Gan, G.; Tang, H.; Cheng, J.; Qi, J.; Ji, J.; Pan, L.; Duan, S.; Wang, W.; Wang, Y.; Cheng, Y.; He, Z.; Su, Z.; Yang, Z.; Pan, Z.; Zeng, A.; Wang, B.; Chen, B.; Shi, B.; Pang, C.; Zhang, C.; Yin, D.; Yang, F.; Chen, G.; Xu, J.; Zhu, J.; Chen, J.; Chen, J.; Chen, J.; Lin, J.; Wang, J.; Chen, J.; Lei, L.; Gong, L.; Pan, L.; Liu, M.; Xu, M.; Zhang, M.; Zheng, Q.; Yang, S.; Zhong, S.; Huang, S.; Zhao, S.; Xue, S.; Tu, S.; Meng, S.; Zhang, T.; Luo, T.; Hao, T.; Tong, T.; Li, W.; Jia, W.; Liu, X.; Zhang, X.; Lyu, X.; Fan, X.; Huang, X.; Wang, Y.; Xue, Y.; Wang, Y.; Wang, Y.; An, Y.; Du, Y.; Shi, Y.; Huang, Y.; Niu, Y.; Wang, Y.; Yue, Y.; Li, Y.; Zhang, Y.; Wang, Y.; Wang, Y.; Zhang, Y.; Xue, Z.; Hou, Z.; Du, Z.; Wang, Z.; Zhang, P.; Liu, D.; Xu, B.; Li, J.; Huang, M.; Dong, Y.; and Tang, J. 2025. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. arXiv:2507.01006.

Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S. C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; Wang, Z.; Fergus, R.; LeCun, Y.; and Xie, S. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. arXiv:2406.16860.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024a. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wang, W.; Ding, L.; Zeng, M.; Zhou, X.; Shen, L.; Luo, Y.; and Tao, D. 2024b. Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal Large Language Models. arXiv:2408.15556.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; Xu, J.; Chen, K.; Xu, B.; Li, J.; Dong, Y.; Ding, M.; and Tang, J. 2023. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*.

Wiedmann, L.; Zohar, O.; Mahla, A.; Wang, X.; Li, R.; Frere, T.; von Werra, L.; Gosthipaty, A. R.; and Marafioti, A. 2025. FineVision: Open Data Is All You Need. arXiv:2510.17269.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; Xie, Z.; Wu, Y.; Hu, K.; Wang, J.; Sun, Y.; Li, Y.; Piao, Y.; Guan, K.; Liu, A.; Xie, X.; You, Y.; Dong, K.; Yu, X.; Zhang, H.; Zhao, L.; Wang, Y.; and Ruan, C. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv preprint arXiv:2412.10302*.

Yuan, L.; Wang, J.; Sun, H.; Zhang, Y.; and Lin, Y. 2025. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. arXiv:2501.07888.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.; Li, L. D.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022. GLIPv2: Unified Vision-Language Pretraining for Language Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, X.; Li, D.; Liu, B.; Bao, Z.; Zhou, Y.; Yang, B.; Liu, Z.; Zhong, Y.; and Yuan, T. 2025a. Layer-wise Vision Injection with Disentangled Attention for Efficient LVLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang, X.; Li, D.; Liu, B.; Bao, Z.; Zhou, Y.; Yang, B.; Liu, Z.; Zhong, Y.; Zhao, Z.; and Yuan, T. 2025b. HiMix: Reducing Computational Complexity in Large Vision-Language Models. *arXiv preprint arXiv:2501.10318*.