

VL-RouterBench: A Benchmark for Vision–Language Model Routing

Zhehao Huang^{1,†} Baijiong Lin^{2,†,‡} Jingyuan Zhang¹ Jingying Wang¹
 Yuhang Liu¹ Ning Lu³ Tao Li¹ Xiaolin Huang^{1,✉}

¹Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University

²The Hong Kong University of Science and Technology (Guangzhou)

³The Hong Kong University of Science and Technology

{kinght.h, tonyzhang666, highmore, yuhangliu, li.tao, xiaolinhuang}@sjtu.edu.cn

{blin241}@connect.hkust-gz.edu.cn {ning.lu}@connect.ust.hk

Abstract

Multi-model routing has evolved from an engineering technique into essential infrastructure, yet existing work lacks a systematic, reproducible benchmark for evaluating vision–language models (VLMs). We present **VL-RouterBench** to assess the overall capability of VLM routing systems systematically. The benchmark is grounded in raw inference and scoring logs from VLMs and constructs quality and cost matrices over sample–model pairs. In scale, VL-RouterBench covers 14 datasets across 3 task groups, totaling 30,540 samples, and includes 15 open-source models and 2 API models, yielding 519,180 sample–model pairs and a total input–output token volume of 34,494,977. The evaluation protocol jointly measures average accuracy, average cost, and throughput, and builds a ranking score from the harmonic mean of normalized cost and accuracy to enable comparison across router configurations and cost budgets. On this benchmark, we evaluate 10 routing methods and baselines and observe a significant routability gain, while the best current routers still show a clear gap to the ideal Oracle, indicating considerable room for improvement in router architecture through finer visual cues and modeling of textual structure. We open-source the complete data construction and evaluation toolchain to promote comparability, reproducibility, and practical deployment in multimodal routing research. Project is available at [VL-RouterBench](#).

1. Introduction

In multi-model systems [1], the router has gradually evolved from an engineering optimization technique into critical infrastructure. As model families diversify and differ markedly in inference cost and capability, a single model can no longer ensure both performance and efficiency across all request

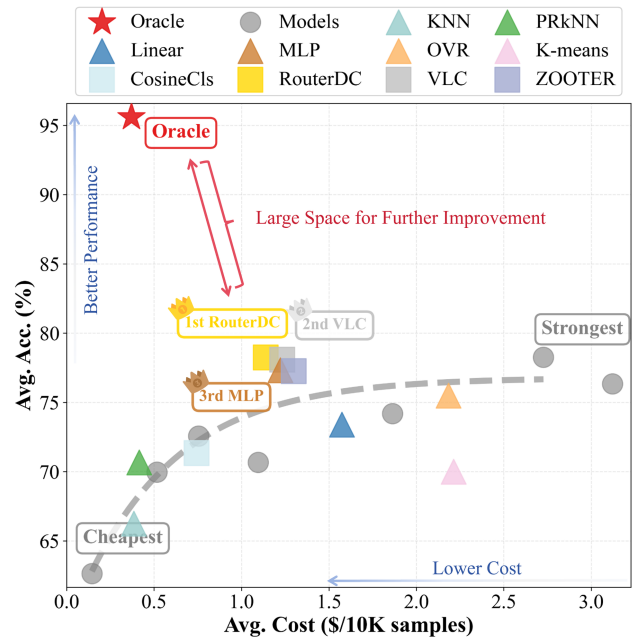


Figure 1. Overall performance comparison on VL-RouterBench. The x and y-axis are Average Cost and Average Accuracy, respectively. Gray dots denote the performance of single models at different costs, and “Strongest” and “Cheapest” mark the baselines that use only the strongest or the cheapest model. The gray dashed curve depicts the Pareto frontier fitted from these single-model points (only models near the frontier are shown for clarity). “1st RouterDC”, “2nd VLC”, and “3rd MLP” indicate the top three routers by Rank Score. Points closer to the upper left reflect a better accuracy–cost trade-off. The results show that even advanced routers still have a noticeable performance gap to the Oracle in the VLM routing setting.

types [2]. To address this, a router [3] performs dynamic model selection at the query level, enabling flexible trade-offs between quality and cost. The routing ecosystem built around this goal is expanding rapidly. In recent years research [4–7] has shown that selecting models at the task or sample level can substantially reduce average inference cost and end-to-end latency while largely preserving output quality, and on this basis more systematic evaluation practices

[†]Equal contribution [‡]Project lead [✉]Corresponding author

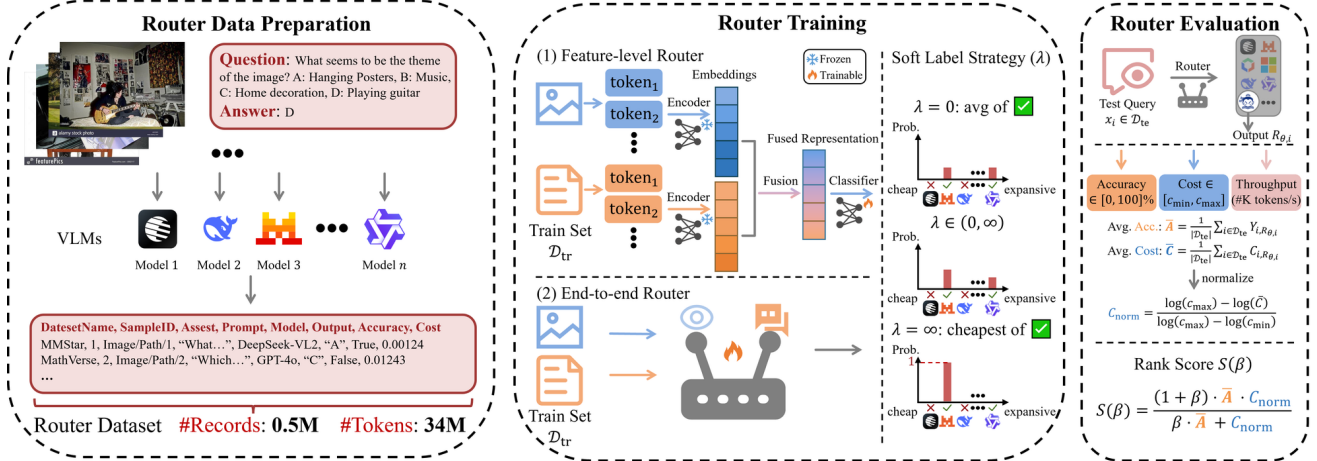


Figure 2. We propose VL-RouterBench to systematically assess the overall performance of vision-language model routing strategies. The three diagrams on the left, middle, and right represent Router Data Preparation (Sec. 3.2), Router Training (Sec. 3.3), and Router Evaluation (Sec. 3.4), respectively.

and open benchmarks [8–11] have emerged, advancing the routing paradigm from static rule based configuration to data driven learned router. At the same time, several mainstream commercial model families have embedded routing capabilities within unified product interfaces so that the underlying system can automatically schedule across models, for example the GPT-5¹ includes built in routing. Construction and evaluation of effective routers has become a shared need for engineering deployment and academic research.

Although routing research for large language models (LLMs) [12] is maturing, the vision-language model (VLM) [13] domain has long lacked a systematic and reproducible routing benchmark. Unlike LLM routing, VLM routing faces more challenges on multiple fronts. First, the task types supported by VLMs are highly diverse, including visual question answering (VQA) [14–18], visual reasoning [19–22], and chart OCR [23–26], and different tasks emphasize different aspects of model capability, which makes it more difficult to define what constitutes an “optimal” routing decision within a unified framework. Second, the mechanism of multimodal fusion remains an open problem [27]. Different VLMs vary substantially in modality interaction and semantic representation, which further increases the difficulty of router design and evaluation. Beyond these challenges, VLM routing also addresses issues introduced by the visual modality, such as the density of visual semantics [28] and cross-modal alignment [29]. These factors collectively distinguish VLM routing from LLM routing and constitute its unique difficulties. In view of these differences, we consider it both necessary and urgent to construct a VLM specific routing benchmark that can drive reproducible progress in the research and deployment of VLM routing systems.

We present **VL-RouterBench** to systematically evaluate the overall capability of VLM routing strategies. To ensure

reproducibility and comparability, we design a complete pipeline around router data preparation, router training, and router evaluation, as shown in Fig. 2. First, for data, we use raw inference and scoring logs from VLMs to construct quality and cost matrices. For each sample we retain the image path and textual instruction, the raw answers produced by each candidate model, correctness labels obtained by rule based evaluation as quality, and the counts of input and output tokens, from which we compute the sample level inference cost. In router training, we employ an adjustable soft label strategy that explicitly controls the trade-off between accuracy and cost, and we consider two routing architectural paradigms, namely feature-level routers and end-to-end routers. In router evaluation, we measure average accuracy, average cost, and throughput for each router on the test set, and we combine normalized cost and average accuracy by a harmonic mean to yield a single rank score, which allows researchers to compare different designs under a unified scale for accuracy and cost. Through large scale data collection, VL-RouterBench covers 14 datasets across three task groups with 30,540 samples, includes 15 open source models and 2 API models, and yields 519,180 sample-model pairs with a total input-output token volume of 34,494,977.

On VL-RouterBench, we systematically evaluate 10 routing methods and baselines, and further analyze the impact of different text and visual encoders, multimodal feature fusion methods, and multimodal backbones on routing performance, yielding conclusions that are consistent across settings and practically informative. First, we observe a significant routability reward. Under comparable or even lower cost, learned routing systems generally deliver stable gains in accuracy over any single model, which indicates that differences in difficulty and structure among multimodal samples can be effectively captured by lightweight routers. Second, text and visual embeddings followed by simple fu-

¹<https://openai.com/index/introducing-gpt-5>

sion schemes, such as normalized concatenation, suffices to support highly competitive routers and is consistently superior to counterparts that use only a single modality, which shows that multimodal input provides additional and fine grained discriminative signals for routing decisions. Finally, there remains a noticeable gap between the best current routing methods and the ideal Oracle, which suggests substantial potential for improving router architectures through finer visual cues and through modeling of layout and textual structure, as shown in **Fig. 1**.

Our main contributions are summarized below:

1. We propose VL-RouterBench, a large scale benchmark for VLM routing systems that provides unified data and protocols for comparison of different routing methods.
2. We design a complete pipeline that covers data preparation, router training, and evaluation, and we introduce an adjustable soft label strategy that enables controllable trade-offs between accuracy and cost during training.
3. We conduct comprehensive comparisons and ablations of multiple routing methods on the benchmark, revealing the performance gap between current methods and the ideal Oracle, and offering guidance for future architectural design of VLM routers.

2. Related Work

2.1. Routing Systems

Multi-model routing systems [1] aim to dynamically select the most suitable model for each input to balance quality, cost, and throughput. Early methods [30], employed heuristic cascading strategies, invoking models in cost order to significantly reduce inference overhead. Later, systems [31] modeled routing as the “quality-cost-throughput” trilemma, validating the comprehensive advantages of learned routing in real-world scenarios. In terms of training methods, existing work can be divided into three categories: supervised learning based on performance logs [8], preference-based binary routing [32], and soft label training based on reward models [4, 33]. Although MoE models [34] also have routing mechanisms, they focus on conditional computations within a single network, which fundamentally differs from routing at the system level. Current research is mainly focused on LLM routers, and this paper extends the routing paradigm to vision–language scenarios, combining feature-level routing with end-to-end encoders to explore the impact of cross-modal representations on routing performance.

2.2. Vision-Language Models

The rapid development of general VLMs [13] provides a rich pool of candidate models for routing. BLIP-2 [35] provides a solution for cost-effective multimodal modeling by using a lightweight query Transformer, leveraging frozen image encoders and large frozen language models for vision-language

pretraining. Building upon this, LLaVA [36] introduces visual instruction fine-tuning. InstructBLIP [37] systematically investigates cross-task instruction fine-tuning processes and module designs, reflecting the impact of training paradigms on generalization. Recently, the Qwen2.5-VL series [27] introduced a dynamic resolution mechanism, assigning adaptive numbers of visual tokens to each image, thereby clearly revealing the trade-off between visual fidelity and computation. In terms of evaluation, VLMEvalKit [38] integrates hundreds of proprietary and open-source large VLMs along with dozens of benchmark datasets under a unified generative protocol, enabling consistent cross-model comparison and large-scale log collection. Based on this ecosystem, our benchmark collects datasets from major VLM task groups, and pairs them with VLMs to provide standardized inference pathways for routing data.

2.3. Router Benchmarks

Routing evaluation for LLMs has gradually moved towards standardization in recent years. RouterBench [8] first proposed a systematic routing evaluation framework, providing theoretical analysis and comparisons across various routers based on a unified log format, offering a template and upper-bound estimation protocol for log-driven routing research. Subsequently, RouterEval [9] scaled up and systematically reported the model-level scaling-up phenomenon resulting from increased candidate pools and router capabilities meeting the mark, emphasizing the structural benefits of the routing paradigm relative to the single strongest model. To enhance horizontal comparability, RouterArena [11] built a routing evaluation dataset that covers a wide range of categories and difficulty levels, accompanied by multi-dimensional metrics including accuracy, cost ratio, optimal selection rate, latency, and robustness, along with an automated leaderboard update process, significantly lowering the barrier for new routing methods to be incorporated and continuously evaluated. All of the above benchmarks focus on text routing, whereas this paper is the first to build a unified evaluation benchmark tailored to the multimodal characteristics of VLMs, providing reproducible training and evaluation protocols.

3. VL-RouterBench Pipeline

3.1. Problem Formulation

We consider a routing framework for VLMs. Let the sample set be $\mathcal{D} = \{x_i = (I_i, T_i)\}_{i=1}^N$, where I_i denotes an image or a set of images and T_i denotes a textual instruction. The candidate multimodal model set is $\mathcal{M} = \{m_j\}_{j=1}^M$. For any sample i , the input is the multimodal pair $x_i = (I_i, T_i)$. A router parameterized by θ takes x_i as input and produces a decision distribution over models $\pi_\theta(\cdot | x_i)$, and during testing the model with the highest probability is selected as

the routed model:

$$R_{\theta,i} = R(\theta, x_i) = \arg \max_{j \in \{1, \dots, M\}} \pi_{\theta}(m_j | x_i). \quad (1)$$

This formalization enables routers to adaptively assign each input, based on its visual and linguistic content, to the most suitable VLM.

To systematically balance routing accuracy and cost, we model router training as a multi objective optimization problem. Let $Y_{i,j}$ and $C_{i,j}$ for each sample–model pair (i, j) denote a performance indicator (such as accuracy) and an inference cost (such as spending dollars). The training objective is

$$\min_{\theta} \underbrace{\mathbb{E}[-Y_{i,R_{\theta,i}}]}_{\text{performance risk}} + \lambda \underbrace{\mathbb{E}[C_{i,R_{\theta,i}}]}_{\text{expected cost}}, \quad (2)$$

where $\lambda \geq 0$ controls the penalty strength on cost. The objective jointly minimizes performance risk and expected cost to achieve a controllable balance of overall effectiveness, which provides a clear theoretical foundation for subsequent router training methods and supports systematic comparison and analysis of different routing strategies.

3.2. Router Data Preparation

We uniformly extract routing data from the inference and scoring artifacts of VLMEvalKit [38] to construct a standardized benchmark that ensures fairness and reproducibility of evaluation. To maintain consistency in quality assessment, we adopt a widely used setting in VLM evaluation where each sample prompt contains a single image and correctness can be verified by rule-based criteria, for example multiple choice and answer matching, thereby avoiding biases introduced by multi image inputs or subjective judgment. During data construction we do not consider task group labels of samples. Instead we organize the minimal data unit that records dataset name, sample index, image path, textual prompt, model name, model output, a boolean label for accuracy, and cost of input and output tokens. Using an automated verification procedure that compares model answers with the ground truth, we obtain a correctness label $Y_{i,j} \in \{0, 1\}$ for each sample–model pair (i, j) and aggregate all records to form the global quality matrix $Y \in \mathbb{R}^{N \times M}$ for accuracy assessment. Cost is computed from token statistics in actual inference logs together with public model price sheets. Specifically, for sample i and model j the cost is

$$C_{i,j} = n_{i,j}^{\text{in}} \cdot c_j^{\text{in}} + n_{i,j}^{\text{out}} \cdot c_j^{\text{out}}, \quad (3)$$

where $n_{i,j}^{\text{in}}$ and $n_{i,j}^{\text{out}}$ denote the numbers of input and output tokens respectively, and c_j^{in} and c_j^{out} are the corresponding prices per million tokens for that model. This yields the cost matrix $C \in \mathbb{R}^{N \times M}$ for cost assessment. VL-RouterBench

covers 14 datasets with 30,540 samples and involves 15 open source models and 2 API models, resulting in 519,180 complete inference records and a total token count of 34,494,977. The specific datasets and model choices are detailed in Sec. 4.1 and Sec. 4.2.

3.3. Router Training

We divide the entire routing dataset into a training set $\mathcal{D}_{\text{tr}}$, a validation set $\mathcal{D}_{\text{dev}}$, and a test set $\mathcal{D}_{\text{te}}$ with the ratio $|\mathcal{D}_{\text{tr}}| : |\mathcal{D}_{\text{dev}}| : |\mathcal{D}_{\text{te}}| = 7 : 1 : 2$. This split is designed to ensure sufficient model training, reliable hyperparameter tuning, and independence of the final evaluation. The validation set is used to optimize the router hyperparameters under preset evaluation metrics so as to strike a reasonable balance between accuracy and cost.

Accuracy–Cost–Aware Soft Label Strategy. To introduce the trade-off between accuracy and cost into router training, we solve the Lagrangian optimization problem in Eq. (2) and obtain an analytic soft label target as

$$t_i^{(\lambda)}(j) = \frac{\mathbf{1}\{Y_{i,j} = 1\} \cdot g_{\lambda}(C_{i,j})}{\sum_{j: Y_{i,j}=1} g_{\lambda}(C_{i,j})}, \quad (4)$$

where $\mathbf{1}(\cdot)$ is the indicator function and $g_{\lambda}(\cdot)$ is a positive function that is monotonically decreasing in cost. Its proof is provided in Appendix D. Intuitively, the router should place probability mass only on models that answer correctly and should prefer a correct model with lower cost. In practice, we adopt an exponential cost decay $g_{\lambda}(c) = \exp(-\lambda c)$, which yields a soft label target whose accuracy–cost trade-off is controlled by the hyperparameter $\lambda \geq 0$:

$$t_i^{(\lambda)}(j) = \frac{\mathbf{1}\{Y_{i,j} = 1\} \cdot \exp(-\lambda \cdot C_{i,j})}{\sum_{j: Y_{i,j}=1} \exp(-\lambda \cdot C_{i,j})}. \quad (5)$$

As shown in Fig. 2, when $\lambda = 0$, cost is ignored and probability is distributed uniformly over all correct models, so the router focuses solely on accuracy. As $\lambda \rightarrow +\infty$, the objective strongly favors low cost, and the soft label degenerates to selecting only the cheapest correct model. For $0 < \lambda < +\infty$, the objective continuously balances accuracy and cost. We then train the router parameters θ with the following soft label cross entropy loss:

$$\mathcal{L}_{\text{soft}}(\theta; \lambda) = \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{i \in \mathcal{D}_{\text{tr}}} \text{KL}\left(t_i^{(\lambda)} \parallel \pi_{\theta}(\cdot | x_i)\right) \quad (6)$$

$$= \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{i \in \mathcal{D}_{\text{tr}}} \sum_j t_i^{(\lambda)}(j) (-\log \pi_{\theta}(m_j | x_i)). \quad (7)$$

Regarding router architectures, we explore two paradigms. The first uses pretrained encoders to extract text and image features separately and then makes routing decisions with a classifier, as the feature-level router. The second processes

multimodal inputs in an end-to-end manner and directly outputs the model selection, as the end-to-end router. Specific implementations are detailed in **Sec. 4.3** and Appendix H.

3.4. Router Evaluation & Metrics

To objectively compare different routers, we establish a multi-dimensional evaluation protocol centered on accuracy, cost, and efficiency. Let $R_{\theta,i}$ denote the final model decision made by the router for sample x_i . (1) **Average Accuracy (Avg. Acc.)** is

$$\bar{A} = \frac{1}{|\mathcal{D}_{te}|} \sum_{i \in \mathcal{D}_{te}} Y_{i,R_{\theta,i}}, \quad (8)$$

measured in percent and reflecting the overall correctness of routing decisions. (2) **Average Cost (Avg. Cost)** is

$$\bar{C} = \frac{1}{|\mathcal{D}_{te}|} \sum_{i \in \mathcal{D}_{te}} C_{i,R_{\theta,i}}, \quad (9)$$

by default presented as \$/10K samples, which quantifies the economic overhead during inference. (3) To promote comparability across router configurations, we follow [11] and introduce a multi-objective **Rank Score** that harmonically averages log normalized cost and accuracy. We first apply bounded normalization to cost. Given the full-sample cost range $[c_{min}, c_{max}]$, the normalized cost is

$$C_{norm} = 100 \times \frac{\log_2(c_{max}) - \log_2(\bar{C})}{\log_2(c_{max}) - \log_2(c_{min})}, \quad (10)$$

which is multiplied by 100 to align with the percentage scale of Avg. Acc. We then combine accuracy and normalized cost using the harmonic mean as

$$S(\beta) = \frac{(1 + \beta) \cdot \bar{A} \cdot C_{norm}}{\beta \cdot \bar{A} + C_{norm}}, \quad (11)$$

where $\beta > 0$ controls the preference between quality and cost. By default we set $\beta = 0.1$ to prioritize accuracy. (4) **Throughput** measures system efficiency as the number of tokens processed per second in thousands of tokens per second (#K tokens/s), which evaluates the response efficiency of a deployed router system. (5) In addition to the quantitative metrics mentioned above, we further describe the overall trade-offs at different operating points by plotting accuracy–cost curves. For each given router, all its operating points in the accuracy–cost plane can be fitted into a curve, forming the **Pareto frontier**. Intuitively, the Pareto frontier closer to the upper left indicates that higher accuracy can be achieved at the same cost, or that lower cost can be achieved at the same accuracy. The detailed process of how to fit the Pareto frontier will be explained in Appendix E.

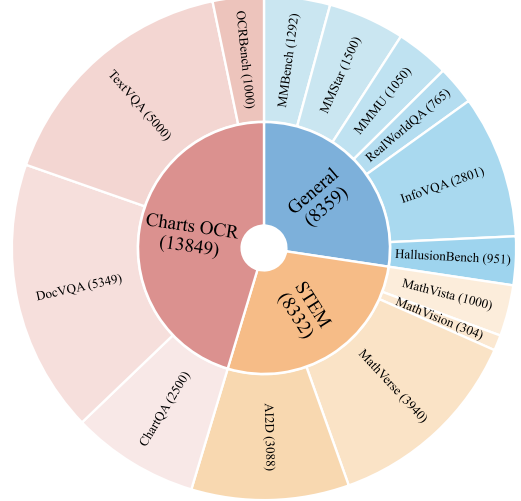


Figure 3. Dataset distribution in VL-RouterBench. The inner ring shows the three groups, and the outer ring lists the individual datasets. The numbers in parentheses indicate the number of samples contained in each entry.

4. Benchmark Curation

4.1. Datasets

Our benchmark curates 14 datasets across three task groups, namely *General*, *STEM*, and *Charts OCR*, in order to induce sufficient routability differences and to cover diverse real application scenarios. The General group includes MM-Bench [14], MMStar [15], MMMU [16], RealWorldQA [17], InfoVQA [18], and HallusionBench [39], which together balance broad knowledge, evaluate cross domain understanding, assess infographic comprehension, and measure robustness. The STEM group covers MathVista [19], MathVision [20], MathVerse [21], and AI2D [22], focusing on structured visual cues and symbolic reasoning in mathematics and scientific diagrams. The Charts and OCR/Document group contains ChartQA [23], DocVQA [24], TextVQA [25], and OCRBench [26], and it systematically examines chart reading and integrated OCR capability. The data distribution is shown in **Fig. 3**, with detailed descriptions of the datasets available in Appendix F. This grouping and unified processing provide a solid foundation for routers to learn generalizable selection boundaries across multiple tasks.

4.2. Models

The core principle in building our model candidate pool is to ensure sufficient heterogeneity across models in terms of scale, architecture, and pricing, in order to realistically simulate the “quality–cost–latency” trilemma faced in actual deployment scenarios. To achieve this, we selected 15 open-source models and 2 API models, covering a parameter range from approximately 1B to 78B parameters, and incorporating a variety of mainstream vision encoding and modality alignment paradigms. The selected open-source

Table 1. VLMs used in our benchmark, sorted by parameter count, with per-million-token prices for input and output. For the MoE model, m -A- n represents the total number of m B parameters in the model and the number of n B parameters activated during inference. All model prices are estimated based on the model size, compared with the prices on `Together.ai`.

Model	Params (B)	Input Price (\$/1M tokens)	Output Price (\$/1M tokens)
Janus-Pro-1B [40]	1.0	0.05	0.05
DeepSeek-VL2-Tiny [41]	27.0-A-1.0	0.05	0.05
SmolVLM2 [42]	2.2	0.06	0.06
Kimi-VL-A3B-Thinking-2506 [43]	16.0-A-2.8	0.20	0.25
Phi-3.5-Vision [44]	4.2	0.10	0.10
DeepSeek-VL2 [41]	27.0-A-4.5	0.35	0.50
Janus-Pro-7B [40]	7.0	0.18	0.25
MiMo-VL-7B-RL [45]	7.0	0.20	0.30
LLaVA-Next-Vicuna-7B [36]	7.0	0.20	0.20
Qianfan-VL-8B [46]	8.0	0.18	0.25
Pixtral-12B [47]	12.0	0.25	0.35
Gemma3-27B [48]	27.0	0.35	0.50
Qwen2.5-VL-32B-Instruct [27]	32.0	0.40	0.60
Qwen2.5-VL-72B-Instruct [27]	72.0	0.80	1.20
InternVL2.5-78B [49]	78.0	1.00	1.50
Gemini-Flash-2.5 [50]	-	0.30	2.40
GPT-4o [51]	-	2.50	10.00

models include: DeepSeek-VL2 [41], DeepSeek-VL2-Tiny [41], Gemma3-27B [48], InternVL2.5-78B [49], Janus-Pro-1B [40], Janus-Pro-7B [40], Kimi-VL-A3B-Thinking-2506 [43], LLaVA-Next-Vicuna-7B [36], MiMo-VL-7B-RL [45], Phi-3.5-Vision [44], Pixtral-12B [47], Qianfan-VL-8B [46], Qwen2.5-VL-32B-Instruct [27], Qwen2.5-VL-72B-Instruct [27], and SmolVLM2 [42]. The API models include GPT-4o [51] and Gemini-Flash-2.5 [50]. Model pricing is derived from the publicly available pricing policies of the AI cloud platform `Together.ai`², supplemented by estimates based on model size for those not listed. This approach ensures that cost metrics are both comparable and additive across all candidate models. A summary of model parameters and associated pricing is provided in **Tab. 1**, with comprehensive model descriptions available in Appendix G.

4.3. Routing Methods

Training-Free Baselines. We begin by introducing a set of training-free routing strategies to establish essential performance references for evaluating subsequent data-driven routers. Among them, **Oracle** establishes the theoretical performance upper bound for any router. It operates with perfect foresight: for each input sample, it selects the lowest-cost model that produces a correct response. If no model can answer the sample correctly, it falls back to the globally cheapest model to minimize cost. Formally, for a given sample x_i , we define the set of correct models as $\mathcal{S}_i \triangleq \{j \in \{1, \dots, M\} \mid Y_{i,j} = 1\}$. The Oracle routing decision is given by:

$$R^*(x_i) = \begin{cases} \arg \min_{j \in \mathcal{S}_i} C_{i,j}, & \text{if } \mathcal{S}_i \neq \emptyset \\ \arg \min_{j \in \{1, \dots, M\}} C_{i,j}, & \text{if } \mathcal{S}_i = \emptyset \end{cases} \quad (12)$$

²<https://www.together.ai/>

Strongest always selects the single model with the highest average accuracy on the entire training set $\mathcal{D}_{\text{tr}}$:

$$R^S(x_i) = \arg \max_{j \in \{1, \dots, M\}} \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{i \in \mathcal{D}_{\text{tr}}} Y_{i,j}. \quad (13)$$

Conversely, **Cheapest** always selects the single model with the lowest average cost over the training set $\mathcal{D}_{\text{tr}}$:

$$R^C(x_i) = \arg \min_{j \in \{1, \dots, M\}} \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{i \in \mathcal{D}_{\text{tr}}} C_{i,j}. \quad (14)$$

We now proceed to describe the two learning-based routing paradigms systematically studied in this benchmark.

Feature-level Routers. Feature-level routers extract embeddings $e_i^T = E^T(T_i)$, $e_i^I = E^I(I_i)$ from the frozen text encoder E^T and visual encoder E^I , respectively. The features from both modalities are then aggregated using a fusion function $F(\cdot, \cdot)$ to form a single feature $z_i = F(e_i^T, e_i^I)$. A lightweight classifier is then trained on this feature to predict the model selection distribution $\pi_\theta(\cdot \mid x_i) = \text{Cls}(z_i)$. We explore various classifier architectures, including **KNN** [52], **PRkNN** [53], **OVR** [54], **K-means** [52], **Linear** [52], and **MLP** [52], among others (see Appendix H for details). The training objective uses the previously defined accuracy-cost-aware soft label cross-entropy loss. The advantage of this paradigm lies in its excellent portability and training stability, which is well-suited for rapid iteration and feature-level ablation studies on large-scale offline logs, such as replacing visual encoders or adjusting fusion strategies, to better understand the contribution of multimodal representations to routability. Corresponding ablation results are presented and analyzed in **Sec. 5.4**.

End-to-End Routers. directly learn the routing decision function from raw images and texts, fine-tuning the entire router network to adapt to specific task combinations. The goal is to capture finer-grained vision-language routing signals on high-difficulty samples. We refer to the design of BERT-Router [32] in traditional language model routing and explore four vision-language encoders as the backbone for routing: VisualBERT [55], UNITER [56], LXMERT [57], and ViLBERT [58], constructing a Vision-Language-Classification (VLC) Router and fine-tuning all parameters using the previously mentioned soft label cross-entropy loss. Additionally, we introduce various existing end-to-end routing training strategies, such as **CosineCls** [4], **RouterDC** [4], **ZOOTER** [33], etc., and evaluate their generalization capabilities in vision-language scenarios. Detailed descriptions of the vision-language encoder structures and routing training strategies can be found in Appendix H, with related architecture ablation studies presented in **Sec. 5.4**.

Table 2. Performance of Router Methods on VL-RouterBench. The best and second-best results except for Oracle are highlighted in **bold** and underlined, respectively. **Avg. Acc.** is average accuracy (%), **Avg. Cost** is average cost (\$/10K samples), **Rank** is sorted by Rank Score, and **Throughput** is measured in #K tokens/s.

Router	Avg. Acc. $\uparrow$	Avg. Cost $\downarrow$	Rank Score $\uparrow$	Rank $\downarrow$	Throughput $\uparrow$
<i>Training-free Baselines</i>					
Oracle	95.60	0.37	93.68	0	-
Strongest	<u>78.01</u>	2.72	68.88	8	-
Cheapest	62.43	0.14	64.63	11	-
<i>Feature-level Routers</i>					
KNN [52]	66.26	<u>0.38</u>	67.13	9	3.14
PRkNN [53]	70.68	0.41	71.09	5	2.73
OVR [54]	75.49	2.18	68.92	7	6.90
K-means [52]	70.01	2.21	64.62	12	31.43
Linear [52]	68.32 $\pm$ 7.45	1.32 $\pm$ 0.33	65.68 $\pm$ 5.69	10	150.74
MLP [52]	77.49 $\pm$ 0.56	1.13 $\pm$ 0.13	74.23 $\pm$ 0.22	3	<u>146.71</u>
<i>End-to-End Routers</i>					
CosineCls [4]	70.24 $\pm$ 0.62	0.61 $\pm$ 0.16	69.90 $\pm$ 0.34	6	7.58
RouterDC [4]	77.52 $\pm$ 1.04	1.04 $\pm$ 0.26	74.59 $\pm$ 1.05	1	6.31
ZOOTER [33]	74.65 $\pm$ 1.34	0.93 $\pm$ 0.15	72.58 $\pm$ 1.03	4	7.25
VLC [32]	78.09 $\pm$ 1.17	1.23 $\pm$ 0.03	<u>74.33</u> $\pm$ 0.51	<u>2</u>	6.74

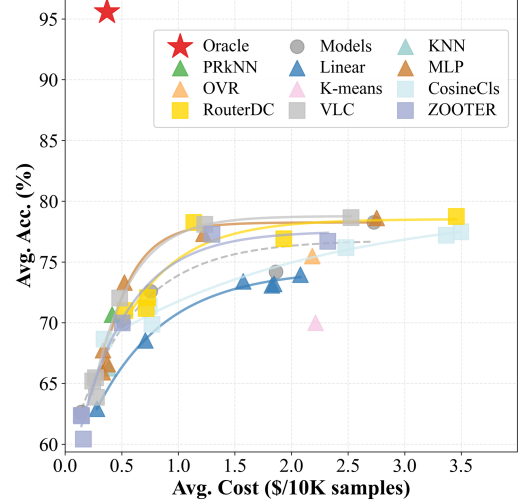


Figure 4. Performance in accuracy-cost plane. Gray dots represent baselines of individual models. Colored markers represent operating points of the same router under different λ , and curves are Pareto fronts fitted to these points. The closer to the upper left, the better the accuracy-cost trade-off.

5. Experiments

5.1. Pilot Study between Oracle and Models

We first evaluate all candidate models along with the Oracle router on VL-RouterBench, and summarize the outcomes in Fig. 1 with full results provided in the Appendix J. The Oracle, representing the ideal routing strategy, achieves the highest accuracy at a very low cost, highlighting the significant potential of routing systems to enhance the cost-effectiveness of VLMs. Moreover, the substantial accuracy gap between the Oracle and any single model indicates considerable variation in model performance across different samples, underscoring the existence of meaningful routing signals that can be exploited. It is also worth noting that certain open-source models deliver higher accuracy at lower costs compared to other candidates, emphasizing the challenge and opportunity in designing routers that effectively balance accuracy and cost in real-world settings.

5.2. Main Results on Router Methods

We evaluate all routing methods on VL-RouterBench after training with soft labels that reflect different accuracy-cost trade-offs. Specifically, we consider $\lambda = \{0, 10, 100, 1000, 10000, +\infty\}$, and the detailed training settings together with the hyperparameters of each method are provided in Appendix J. Tab. 2 reports, for each method, the results corresponding to its best Rank Score under the different trade-off settings. Most routing methods surpass the Strongest baseline, which is the single model with the highest accuracy, indicating that routing can indeed achieve high performance at lower cost than any single model. With a multimodal encoder architecture, the improved RouterDC

attains the best Rank Score and reaches accuracy close to the strongest single model while using only one fifth of the cost. Fig. 4 presents the accuracy-cost Pareto frontiers of all methods to assess router performance across the global cost span. RouterDC and MLP significantly outperform the other routing methods and achieve similar Pareto optimal fronts. Nevertheless, there remains a substantial gap to the optimal Oracle, which suggests ample room for developing routers with better accuracy-cost trade-offs, while our benchmark provides a reproducible platform for such experimentation.

5.3. Latency Profile

As the central component for model selection, the router must satisfy low latency and high throughput. The Throughput metric in Tab. 2 shows that end to end router methods usually achieve a better accuracy and cost trade-off than feature level routers, yet their throughput is slightly lower due to the complexity of the backbone itself. For feature level routers the primary throughput bottleneck lies in the extraction of feature embeddings. Our benchmark offers a solid platform for studying lighter and higher throughput multimodal architectures specifically designed for routers.

5.4. Ablation Study

Text and Visual Encoders. We conduct ablations on the encoders used to extract text and visual embeddings in the feature level router. Considering different embedding dimensions and architectures, we evaluate 5 text encoders, namely **E5-small/base/large-v2** [59], **GTE-large** [60], and **BGE-M3** [61], and 5 visual encoders, namely **DINOv2-small/base/large** [62], **CLIP-L-14** [29], and **SigLIP-L**

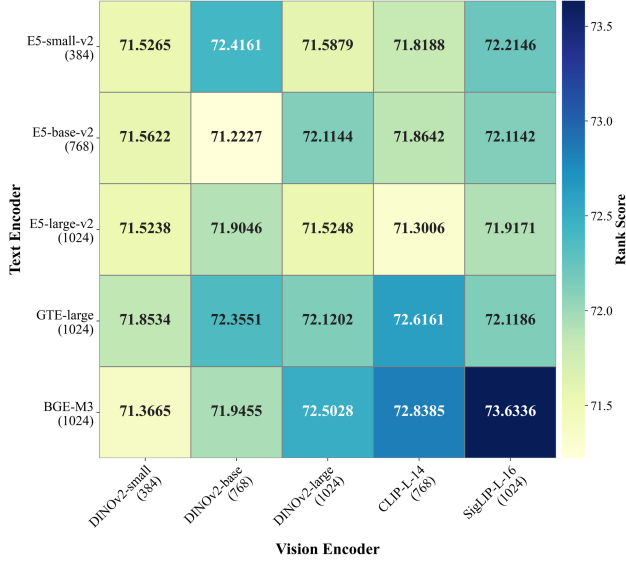


Figure 5. Performance comparison of different text and visual encoders paired on MLP with $\lambda = 100$. The brackets indicate the dimension of the encoder’s output embedding. Other settings are the same as in Table 2.

16 [63]. We pair them in all combinations and train an MLP router. As shown in Fig. 5, increasing the dimensionality of text and visual embeddings effectively improves the router Rank Score. We adopt the best combination, BGE-M3 and SigLIP-L-16, as the embedding configuration for the feature level router.

Fusion Methods. We further examine fusion strategies after obtaining text and visual embeddings. We consider using only the text embedding (**Only-Text**), only the image embedding (**Only-Image**), simple concatenation (**Concat**), normalization followed by concatenation (**Normalize-Concat**), and weighted fusion (**Weighted-Average**). We also include two methods that introduce additional projection parameters for fine tuning, **GMU** [64] and **MLB** [65]. Definitions of these methods are provided in Appendix H. Using the fused features, we train an MLP and report results in Tab. 3. As expected, Only-Text and Only-Image perform weaker because they use only one modality. GMU and MLB respectively emphasize accuracy and cost, but they are not competitive in terms of Rank Score. In contrast, Normalize-Concat yields the best accuracy and cost trade-off and is adopted as the default fusion method for all feature level routers. This indicates that there remains ample space to explore multimodal embedding fusion methods tailored for routing.

End-to-End Router Backbones. We next study the impact of the backbone in end-to-end routers. In addition to the 4 multimodal encoders that process both text and images in Sec. 4.3, we include two single modality baselines: **BERT** [66] from typical LLM router settings using only text, and **MobileNetV4** [67] from the vision modality using only images. The performance of different backbones is reported

Table 3. Performance comparison of different fusion methods on MLP with $\lambda = 100$. The best and second-best results are highlighted in **bold** and underlined, respectively. Other settings are the same as in Table 2. Results are presented as mean and standard deviation across 5 independent trials.

Fusion Method	Avg. Acc. $\uparrow$	Avg. Cost $\downarrow$	Rank Score $\uparrow$	Throughput $\uparrow$
Only-Text	71.41 $\pm$ 1.42	0.60 $\pm$ 0.11	70.97 $\pm$ 1.14	350.96
Only-Image	71.89 $\pm$ 0.61	<u>0.50$\pm$0.13</u>	71.82 $\pm$ 0.42	<u>250.39</u>
Concat	72.35 $\pm$ 0.93	0.63 $\pm$ 0.06	71.70 $\pm$ 1.03	146.71
Normalize-Concat	74.61 $\pm$ 1.31	0.55 $\pm$ 0.11	74.05$\pm$1.06	145.12
Weighted-Average	<u>74.78$\pm$1.02</u>	0.64 $\pm$ 0.10	<u>73.81$\pm$1.09</u>	146.41
GMU [64]	81.81$\pm$1.71	3.32 $\pm$ 0.09	69.12 $\pm$ 1.44	141.25
MLB [65]	66.43 $\pm$ 0.14	0.30$\pm$0.04	67.66 $\pm$ 0.57	141.04

Table 4. Performance comparison of different backbones on VLC using $\lambda = 100$. The best and second-best results are highlighted in **bold** and underlined, respectively. Other settings are the same as in Table 2. Results are presented as mean and standard deviation across 5 independent trials.

Backbone	Image	Text	Avg. Acc. $\uparrow$	Avg. Cost $\downarrow$	Rank Score $\uparrow$	Throughput $\uparrow$
BERT [66]	$\times$	$\checkmark$	63.15 $\pm$ 0.85	0.30$\pm$0.02	64.55 $\pm$ 0.32	8.10
MobileNetV4 [67]	$\checkmark$	$\times$	76.55$\pm$1.14	3.92 $\pm$ 0.45	63.33 $\pm$ 0.72	8.93
VisualBERT [55]	$\checkmark$	$\checkmark$	70.54 $\pm$ 1.12	0.43 $\pm$ 0.12	<u>70.89$\pm$0.72</u>	6.32
UNITER [56]	$\checkmark$	$\checkmark$	67.69 $\pm$ 1.10	0.39 $\pm$ 0.06	68.43 $\pm$ 0.72	6.44
LXMERT [57]	$\checkmark$	$\checkmark$	<u>71.43$\pm$1.03</u>	0.51 $\pm$ 0.02	71.36$\pm$0.81	<u>6.74</u>
ViLBERT [58]	$\checkmark$	$\checkmark$	70.55 $\pm$ 1.08	0.45 $\pm$ 0.10	70.81 $\pm$ 0.68	5.76

in Tab. 4. Because they consider only one modality, BERT and MobileNetV4 achieve markedly lower Rank Score than the four multimodal encoders. Among the multimodal encoders the results are very close, and we choose LXMERT, which attains the highest Rank Score, as the backbone for the end-to-end router. We expect that our benchmark could facilitate the development of multimodal encoder architectures designed for routing to further improve performance.

6. Conclusion

In this paper, we introduced VL-RouterBench, a systematic and reproducible benchmark for routing in VLM systems. Grounded in raw inference and scoring logs from a pool of 15 open source models and 2 API models over 14 datasets for a total of 519,180 records, the benchmark constructs aligned sample–model quality and cost matrices that capture performance–cost trade-offs at scale and support a unified ranking score. On top of this construction, we defined an evaluation protocol that measures average accuracy, average cost, and throughput, and we evaluated 10 routing methods and baselines. The results reveal a clear routability reward: learned routers consistently outperform any single model at comparable or lower cost, while a sizable gap to the ideal Oracle frontier persists, especially in scenarios that require finer visual cues and more precise modeling of textual structure. We open source the complete pipeline. Looking ahead, VL-RouterBench can be naturally extended to more complex VLM settings, such as multiple image inputs or datasets with subjective or preference-based evaluation, thereby further broadening the application scope of routing and bringing benchmarks closer to real-world multimodal workflows.

Acknowledgment. The research leading to these results has received funding from the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (No. 2025-GZL-RGZN-BTBX-02026) and the National Natural Science Foundation of China (No. 62376155).

References

- [1] Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*, 2024. 1, 3
- [2] Olaide N Oyelade, Hui Wang, and Karen Rafferty. A survey of adaptation of large language models to idea and hypothesis generation: Downstream task adaptation, knowledge distillation approaches and challenges. *ACM Computing Surveys*, 2025. 1
- [3] Yiqun Zhang, Hao Li, Jianhao Chen, Hangfan Zhang, Peng Ye, Lei Bai, and Shuyue Hu. Beyond gpt-5: Making llms cheaper and better via performance-efficiency optimized routing. *arXiv preprint arXiv:2508.12631*, 2025. 1
- [4] Shuhao Chen, Weisen Jiang, Baijiong Lin, James T. Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. In *NeurIPS*, 2024. 1, 3, 6, 7, 5
- [5] Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, Guanhao Zhao, Fei Wang, and Runze Wu. Irt-router: Effective and interpretable multi-llm routing via item response theory. In *ACL (1)*, pages 15629–15644. Association for Computational Linguistics, 2025.
- [6] Zhou Chen, Zhiqiang Wei, Yuqi Bai, Xue Xiong, and Jianmin Wu. Tagrouter: Learning route to llms through tags for open-domain text generation tasks. In *ACL (Findings)*, pages 21539–21564. Association for Computational Linguistics, 2025.
- [7] Tao Feng, Yanzen Shen, and Jiaxuan You. Graphrouter: A graph-based router for LLM selections. In *ICLR*. OpenReview.net, 2025. 1
- [8] Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system, 2024. 2, 3
- [9] Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms, 2025. 3
- [10] Tao Feng, Haozhen Zhang, Zijie Lei, Pengrui Han, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Jiaxuan You. Fusionfactory: Fusing llm capabilities with multi-llm log data, 2025.
- [11] Yifan Lu, Rixin Liu, Jiayi Yuan, Xingqi Cui, Shenrun Zhang, Hongyi Liu, and Jiarong Xing. Routerarena: An open platform for comprehensive comparison of llm routers, 2025. 2, 3, 5
- [12] Andrea Matarazzo and Riccardo Torlone. A survey on large language models with some insights on their capabilities and limitations. *arXiv preprint arXiv:2501.04040*, 2025. 2
- [13] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025. 2, 3
- [14] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2, 5, 3
- [15] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 5, 3
- [16] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5, 3
- [17] xAI. Grok-1.5 vision preview, 2024. 5, 3
- [18] Minesh Mathew, Viraj Bagal, Ruben Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2, 5, 3
- [19] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2, 5, 3
- [20] Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 5, 4
- [21] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 5, 4
- [22] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 2, 5, 4
- [23] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 2, 5, 4
- [24] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 5, 4
- [25] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach.

- Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5, 4
- [26] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 2, 5, 4
- [27] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 6, 5
- [28] Luca Rossetto, Cristina Sarasua, and Abraham Bernstein. Estimating the semantic density of visual media. In *ACM Multimedia*, pages 4601–4609. ACM, 2024. 2
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2, 7
- [30] Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023. 3
- [31] Dimitris Stripelis, Zhaozhuo Xu, Zijian Hu, Alay Dilipbhai Shah, Han Jin, Yuhang Yao, Jipeng Zhang, Tong Zhang, Salman Avestimehr, and Chaoyang He. Tensoropera router: A multi-model router for efficient LLM inference. In *EMNLP (Industry Track)*, pages 452–462. Association for Computational Linguistics, 2024. 3
- [32] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *ICLR*. OpenReview.net, 2025. 3, 6, 7
- [33] Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models. In *NAACL-HLT*, pages 1964–1974. Association for Computational Linguistics, 2024. 3, 6, 7
- [34] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025. 3
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023. 3
- [36] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, May 2024. 3, 6, 4
- [37] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 3
- [38] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACM Multimedia*, pages 11198–11201. ACM, 2024. 3, 4
- [39] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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*, pages 14375–14385, 2024. 5, 3
- [40] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 6, 4
- [41] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6, 4
- [42] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 6, 5
- [43] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 6, 4
- [44] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vad-

- damanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuo-hang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 6, 5
- [45] Xiaomi LLM-Core Team Zihao Yue, Zhenrui Lin, Yi-Hao Song, Weikun Wang, Shu-Qin Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zi-Ang Jiang, Zhixian Zheng, Zhichao Song, Zhen Luo, Yue Yu, Yudong Wang, Yu Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xin dan Xu, Xin Ran Song, Xing Zhang, Xing Yong, Xin Zhang, Xia Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Wei Ji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shi liang Yu, Shao yang Liu, Shan yong Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kang Zhou, Kang Zhou, Kai Fang, Jun-Miao Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongsheng Xu, Hengxu Qu, Hao-Song Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong-Yi Ma, Chang Liu, Can Cai, and Bing Xia. Mimo-vl technical report. *ArXiv*, abs/2506.03569, 2025. 6, 4
- [46] Daxiang Dong, Mingming Zheng, Dong Xu, Bairong Zhuang, Wenyu Zhang, Chunhua Luo, Haoran Wang, Zijian Zhao, Jie Li, Yuxuan Li, et al. Qianfan-vl: Domain-enhanced universal vision-language models. *arXiv preprint arXiv:2509.18189*, 2025. 6, 5
- [47] Praveesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Moncault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 6, 5
- [48] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 6, 4
- [49] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 4
- [50] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6, 5
- [51] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 5
- [52] Tal Shnitzer, Anthony Ou, Mirian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*, 2023. 6, 7, 5
- [53] Yandong Zheng, Hui Zhu, Rongxing Lu, Yunguo Guan, Songnian Zhang, Fengwei Wang, Jun Shao, and Hui Li. Efficient and privacy-preserving aggregated reverse knn query over crowd-sensed data. *IEEE Trans. Inf. Forensics Secur.*, 18:4285–4299, 2023. 6, 7, 5
- [54] Ryan M. Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141, 2004. 6, 7, 5
- [55] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 6, 8, 7
- [56] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 6, 8, 7
- [57] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics, 2019. 6, 8, 7
- [58] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 6, 8, 7
- [59] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 7
- [60] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 7
- [61] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. 7
- [62] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 7
- [63] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11941–11952. IEEE, 2023. 8
- [64] John Arevalo, Tamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. 8, 6
- [65] Kankana Roy. Multimodal score fusion with sparse low-rank bilinear pooling for egocentric hand action recognition. *ACM*

Trans. Multim. Comput. Commun. Appl., 20(7):227:1–227:22, 2024. [8](#), [6](#)

- [66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019. [8](#), [6](#)
- [67] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby R. Banbury, Chengxi Ye, Berkin Akin, Vaibhav Aggarwal, Tenghui Zhu, Daniele Moro, and Andrew G. Howard. Mobilenetv4: Universal models for the mobile ecosystem. In *ECCV (40)*, volume 15098 of *Lecture Notes in Computer Science*, pages 78–96. Springer, 2024. [8](#), [6](#)