VL-ROUTERBENCH

A Benchmark for Vision–Language Model Routing

Zhehao Huang, Baijiong Lin, Jingyuan Zhang, Jingying Wang, Yuhang Liu, Ning Lu, Tao Li, Xiaolin Huang

Shanghai Jiao Tong University | HKUST (Guangzhou) | HKUST

POSTER PRESENTATION

Abstract & Motivation

The Need for VLM Routing

Infrastructure Evolution

Multi-model routing has evolved from an engineering technique into **essential infrastructure** for large-scale AI systems.

Systematic Gap

Existing work lacks a systematic, reproducible benchmark for evaluating **Vision–Language Models (VLMs)** in routing scenarios.

Our Contribution

VL-RouterBench

A large-scale benchmark to assess the overall capability of VLM routing routing systems systematically.

Quality & Cost Matrices

Constructed from **raw inference logs**, covering 14 datasets, 17 models, and over 500,000 sample-model pairs.

Challenges in VLM Routing

Diverse Task Types

VLMs support a wide range of tasks from **VQA** to **Visual Reasoning** and **Chart OCR**. Each task emphasizes different aspects of model capability, making routing difficult to define.

Multimodal Fusion

The mechanism of **multimodal fusion** remains an open problem. Different VLMs vary substantially in modality interaction, increasing the complexity of router design.

Visual Modality Issues

Routing must address issues specific to the visual modality, such as **visual semantic density** and **cross-modal alignment**, which are not present in traditional LLM routing.

VL-RouterBench Pipeline

1. Data Preparation

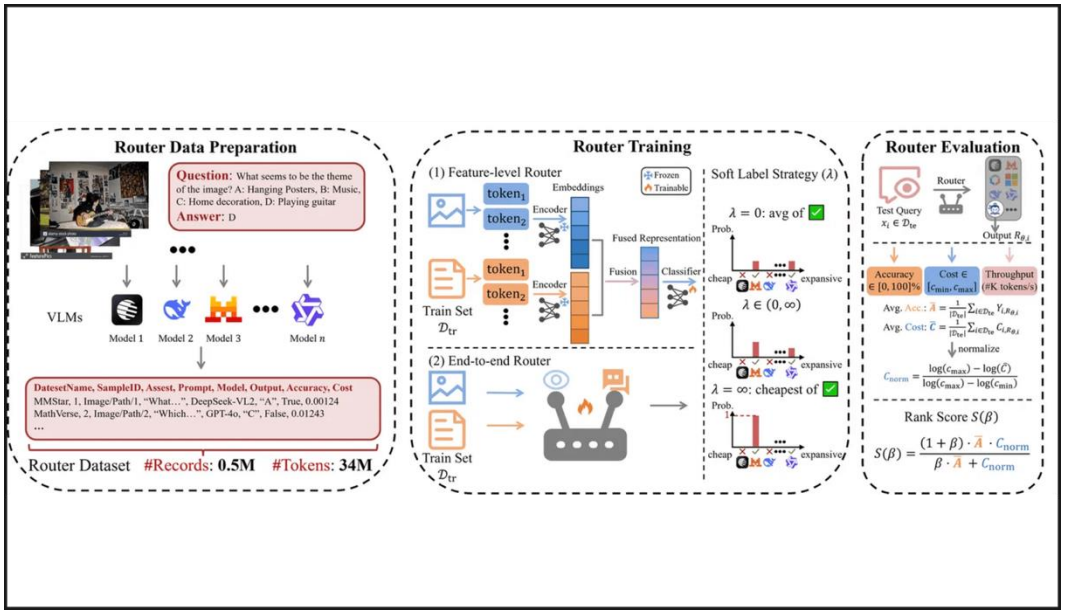
Extracting **raw inference logs** from VLMEvalKit: dataset names, names, prompts, model outputs, accuracy, and cost metrics. metrics.

2. Router Training

Implementing an **adjustable soft label strategy** to control trade-offs between accuracy and inference cost.

3. Router Evaluation

Measuring **accuracy, cost, and throughput** jointly. A harmonic harmonic Rank Score provides comprehensive comparison.



Dataset & Model Distribution

14 Diverse Datasets

General Group

MMBench, MMStar, MMMU, RealWorldQA, InfoVQA, HallusionBench. Focus on **broad knowledge** and robustness.

STEM Group

MathVista, MathVision, MathVerse, AI2D. Focus on **symbolic reasoning** and scientific diagrams.

Charts & OCR Group

ChartQA, DocVQA, TextVQA, OCRBench. Focus on **structured visual cues** and text extraction.

17 Heterogeneous Models

Open-Source Pool (15 Models)

Ranging from **1B to 78B parameters**. Includes DeepSeek-VL2, Qwen2.5-VL, InternVL2.5, Janus-Pro, and LLaVA-Next.

API Models (2 Models)

Industry leaders: **GPT-4o** and **Gemini-Flash-2.5** for high-performance baselines.

Total Samples: 30,540

Sample-Model Pairs: 519,180

Total Tokens: 34,494,977

Routing Methods & Baselines

TRAINING-FREE

Oracle: Theoretical upper bound selecting the lowest-cost correct model.
Strongest/Cheapest: Static baselines prioritizing either accuracy or cost.

FEATURE-LEVEL

Extracts frozen embeddings from text and visual encoders. Uses lightweight classifiers like **MLP, KNN,** and **Linear** models to predict selection distributions.

END-TO-END

Fine-tunes the entire router network (e.g., **VLC Router**) to capture fine-grained routing signals directly from raw multimodal inputs.

Unified Objective

All learning-based methods are trained using **Soft Label Cross-Entropy Loss** to balance performance risk and expected cost.

Main Results & Performance

Significant Routability Reward

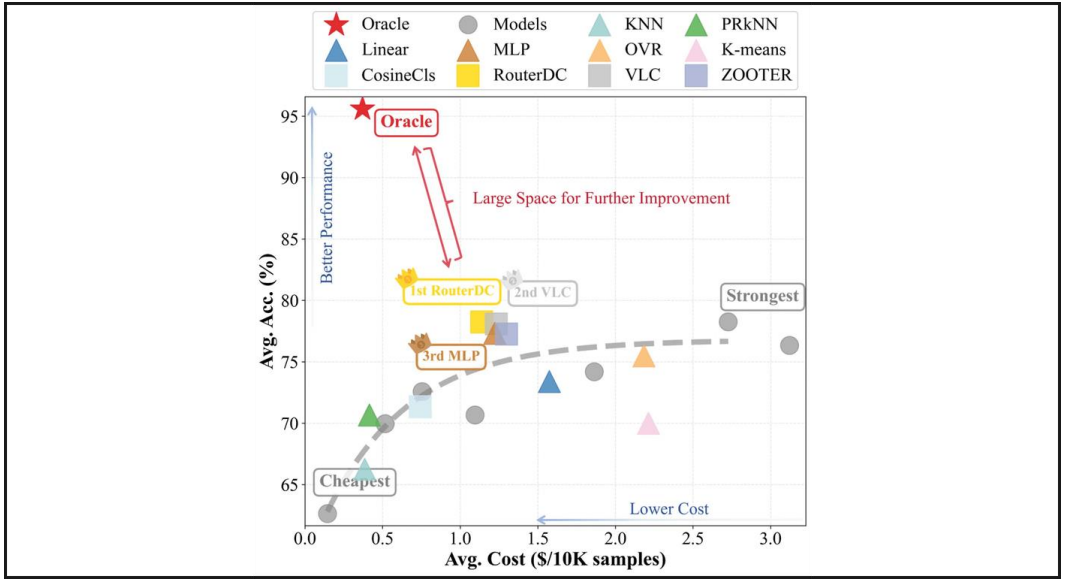
Learned routers consistently deliver stable accuracy gains over over single models at **lower costs**.

Top Performing Routers

RouterDC and **MLP** achieve the best Rank Scores, approaching approaching Strongest accuracy at 1/5th the cost.

Gap to Oracle

A noticeable performance gap still exists between the best current routers and the **ideal Oracle**, indicating room for improvement.



Ablation Study: Modality & Fusion

Multimodal Superiority

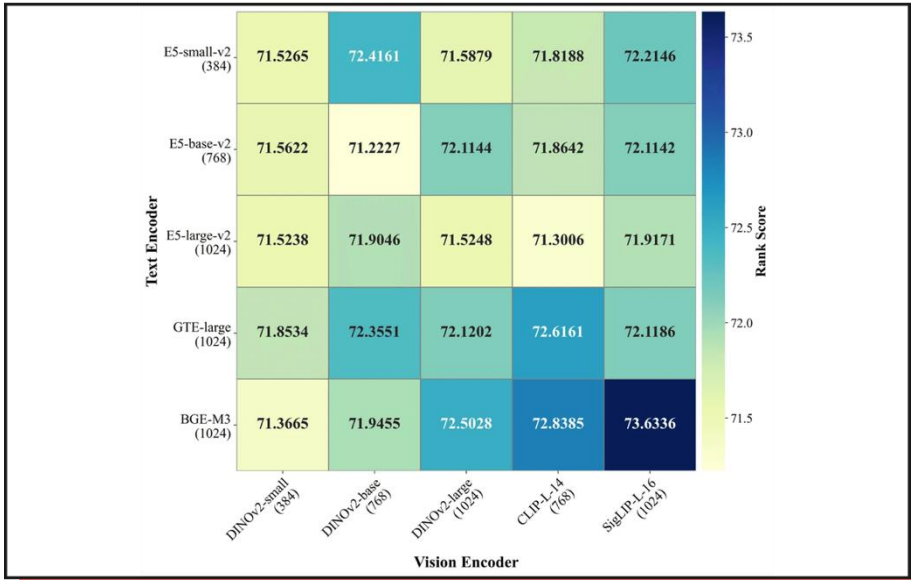
Combining text and visual embeddings provides **better discriminative signals** than single modalities alone.

Fusion Strategies

Normalized concatenation and weighted average schemes are are sufficient for building highly competitive routers.

Encoder Impact

Increasing embedding dimensionality improves Rank Score. Score. Best combo: **BGE-M3** (text) and **SigLIP-L-16** (visual).



Latency & Throughput Analysis

System Efficiency

Real-World Deployment

Routers must keep **ultra-low latency** so model-selection overhead does not negate routing benefits.

Throughput Bottlenecks

Feature-level routers are often limited by **embedding extraction** from encoders.

Deployment Feasibility

End-to-End Trade-offs

End-to-end methods give better accuracy-cost trade-offs but can have **lower throughput**.

Practical Guidance

VL-RouterBench helps study **lighter multimodal architectures** optimized for fast routing.

Conclusion & Future Work

Key Contributions

Unified Foundation

VL-RouterBench provides the first systematic, reproducible benchmark for for VLM routing research and deployment.

Routability Reward

Demonstrated that multimodal routing effectively bridges the gap between performance and cost in large-scale systems.

Future Directions

Architectural Improvements

Exploring finer visual cues and better modeling of textual structure to further further bridge the gap to the ideal Oracle.

Open Sourced

Complete data construction and evaluation toolchain are available to promote comparability and reproducibility in the community.