ReLope: KL-Regularized LoRA Probes for Multimodal LLM Routing

Anonymous ACL submission

Abstract

Routing has emerged as a promising strategy for balancing performance and cost in large language model (LLM) systems that combine lightweight models with powerful but expensive large models. Recent studies show that *probe routing*, which predicts the correctness of a small model using its hidden states, provides an effective solution in text-only LLMs. However, we observe that these probes degrade substantially when applied to multimodal LLMs (MLLMs). Through empirical analysis, we find that the presence of visual inputs weakens the separability of correctness signals in hidden states, making them harder to extract using standard probe designs. To address this challenge, we introduce two complementary approaches for improving probe routing in MLLMs. First, we propose the *Attention Probe*, which aggregates hidden states from the preceding layer based on attention scores to recover distributed correctness signals. Second, we present the *KL-Regularized LoRA Probe (ReLope)*, which inserts a lightweight LoRA adapter and applies a KL regularizer to learn routing-aware representations. Comprehensive experiments demonstrate that our methods consistently outperform strong routing baselines, suggesting that improving the quality of hidden states is key to effective routing in MLLMs.

1 Introduction

Large Language Models (LLMs) have rapidly advanced and are now widely deployed in real-world applications (Zhao et al., 2023). However, their deployment faces a fundamental trade-off: large models provide strong performance but introduce substantial costs, while smaller models are cheaper but less capable. This tension has motivated the development of *routing strategies*, which dynamically decide whether a query can be handled by a lightweight model or should be escalated to a stronger one (Chen et al., 2023, 2025; Jitkrittum et al., 2025). Such strategies enable hybrid LLM systems that achieve favorable trade-offs between performance and efficiency.

Most existing routing methods have been studied in the context of text-only LLMs. Recently, however, multimodal LLMs (MLLMs) (Yin et al., 2024; Liang et al., 2024) have emerged as a new foundation for AI systems capable of jointly processing text and visual inputs. While these models enable more complex reasoning tasks, they also introduce new challenges for efficient deployment. In particular, visual inputs significantly increase inference cost, creating an even stronger need for effective routing mechanisms in multimodal systems.

Among existing strategies, the trained probe has emerged as one of the most effective approaches (Gupta et al., 2024; Chuang et al., 2025). It trains lightweight classifiers on hidden states of the small model to estimate whether the generated answer is correct. Compared with heuristic confidence measures such as token probabilities and average entropy (Huang et al., 2023; Mahaut et al., 2024), probes capture richer semantic signals from LLMs' hidden states, leading to a strong ability in correctness estimation. Because probes operate directly on internal hidden states, they are also architecture-agnostic and, in principle, applicable across different model modalities. However, it remains unclear whether hidden state probes remain reliable when visual inputs are introduced. This naturally raises a research question: *Can probe routings designed for text-only LLMs transfer effectively to MLLMs, and if not, how should they be adapted?*

Through carefully designed experiments, we show that probes effective in text-only LLMs degrade in multimodal settings, achieving consistently lower AUC on queries containing images across multiple datasets. For example, when trained on ScienceQA (Saikh et al., 2022), probe performance drops substantially for questions containing images. This phenomenon is further con-

firmed on A-OKVQA (Schwenk et al., 2022) and MMMU (Yue et al., 2024). These results suggest that the hidden states produced by MLLMs provide weaker signals for predicting correctness.

To address this challenge, we propose two complementary techniques for improving probe routing in MLLMs. First, we introduce the *Attention Probe*, which aggregates token-level hidden states from the preceding layer using an attention mechanism. It recovers distributed correctness signals that may be diluted in the final token hidden state. Second, we propose the *KL-Regularized LoRA Probe (ReLope)*, which adapts hidden representations through a lightweight Low-Rank Adaptation (LoRA) module (Hu et al., 2022) and regularizes them using a KL divergence inspired by the Variational Information Bottleneck (VIB) theory (Alemi et al., 2016). This regularization encourages the model to learn compressed representations that retain information relevant for routing decisions while suppressing task-irrelevant visual noise. Comprehensive experiments show that both Attention Probe and ReLope consistently improve routing performance over existing baselines, with ReLope achieving the strongest trade-offs between accuracy and cost in hybrid MLLM systems.

Our contributions are summarized as follows:

- We identify and empirically characterize a previously underexplored limitation of probe routing: its performance degrades in MLLMs due to weaker correctness signals in hidden representations.
- We propose two effective routing techniques in MLLMs based on the trained probe: the *Attention Probe*, which aggregates token-level hidden states to recover distributed routing signals, and *ReLope*, which learns routing-aware representations via LoRA adaptation and KL regularization.
- Extensive experiments on five multimodal benchmarks demonstrate that our methods consistently outperform existing routing approaches, significantly improving routing quality while maintaining efficient inference in hybrid MLLM systems.

2 Related Work

2.1 Routing Before the Small Model

One family of methods determines routing decisions before querying the small model, relying only on the input queries (Shah and Shridhar, 2025; Ong et al., 2024; Panda et al., 2025; Zhang et al., 2025). Early works train a pretrained router to classify which model will succeed for a given query. Examples include classifier-based routers such as (Shnitzer et al., 2023) and SelectLLM (Ong et al., 2024), reward-based routers like (Lu et al., 2023), and cost-aware strategies like Routoo (Mohammadshahi et al., 2024). Others use non-pretrained strategies, such as Eagle (Zhao et al., 2024), which leverage heuristic selection without extra training.

These methods avoid invoking the small model when unnecessary, reducing cost. However, they are highly dependent on the quality of collected data for training or comparison. More importantly, these methods focus on the text-only prompts. When extending to MLLMs, their router architectures should be redesigned to incorporate additional modalities such as images or videos, which substantially increases implementation complexity.

2.2 Routing With the Small Model

A second class of methods leverages information from the small model. This includes both its final outputs (cascade) and hidden states probe. Cascade approaches run the small model and then apply a rule to decide whether escalation to the large model is needed according to the output of the small model. FrugalGPT (Chen et al., 2023) highlights that cascades can substantially reduce cost while maintaining quality. Most subsequent strategies use different rules to evaluate the small model's confidence as a routing score (Zhao et al., 2024; Chuang et al., 2025; Yue et al., 2023). For example, MoT (Yue et al., 2023) applies multiple inferences and refers to the answer consistency to route. The work (Chuang et al., 2025) evaluates the performance of existing uncertainty estimation methods in the routing problem. Another work (Gupta et al., 2024) shows that sequence-level uncertainty is biased (favoring short or long generations), and proposes refined rules using token-level uncertainty quantiles.

Another line of work directly trains a probe on hidden states of the small model to predict whether its answer will be correct. Post-hoc embedding probe methods (Chuang et al., 2025; Gupta et al., 2024; Mahaut et al., 2024) show that probes trained on intermediate layers can effectively capture correctness signals. These methods have achieved strong performance in text-only routing, balancing efficiency and reliability. However, we empirically

observe that the introduction of images leads to degraded routing accuracy, motivating us to design new routing strategies.

3 Preliminaries

3.1 Problem Definition.

We consider a hybrid MLLM system composed of a small model, such as an on-device model, and a large model that is more capable but also more expensive, typically deployed in the cloud. In this setting, a router is used to determine whether the small model is sufficient for a given query $\mathbf{x}$ or whether the query should be escalated to the large model. Routing can therefore be naturally formulated as correctness prediction for the small model: if the small model is expected to answer correctly, the system uses its output; otherwise, the query is deferred to the large model.

Formally, let $h_s(\mathbf{x})$ denote the prediction of the small model and $h_l(\mathbf{x})$ denote the prediction of the large one. The router $r(\mathbf{x})$ makes a binary decision:

$$r(\mathbf{x}) = \begin{cases} 0, & \text{defer to } h_l(\mathbf{x}), \\ 1, & \text{use } h_s(\mathbf{x}). \end{cases}$$

Let $Y \in \{0, 1\}$ denote whether the output of the small model is correct with respect to the ground truth, where $Y = 1$ indicates correctness. The goal is to design $r(\mathbf{x})$ to accurately predict Y, thereby using the small model when it is likely to be correct and deferring to the large model otherwise.

3.2 Probes Trained on Hidden States.

A representative model-aware routing approach uses hidden states of the small model to predict whether its output $h_s(\mathbf{x})$ will be correct. Previous works have demonstrated that hidden states provide a strong signal of the LLM's confidence, thereby achieving strong performance in text-only LLM routing (Chuang et al., 2025; Gupta et al., 2024; Mahaut et al., 2024).

Formally, let θ denote the parameters of the LLM. Given an input query $\mathbf{x} = (x_1, \ldots, x_n)$ with sequence length n, the model produces a sequence of hidden states at layer l:

$$\mathbf{Z}^{(l)} = \{\mathbf{z}_1^{(l)}, \mathbf{z}_2^{(l)}, \ldots, \mathbf{z}_n^{(l)}\} \in \mathbb{R}^{n \times d},$$

where d denotes the hidden dimension. Following previous work (Chuang et al., 2025; Mahaut et al., 2024), the last-token hidden state $\mathbf{z}_n^{(l)}$ is used as the input of the probe.

A probe $g_\phi : \mathbb{R}^d \to [0, 1]$, parameterized by ϕ, is trained to predict the correctness label Y. The prediction is defined as

$$\hat{y} = \sigma\big(g_\phi(\mathbf{z}_n^{(l)})\big),$$

where σ denotes the logistic sigmoid function. The probe is trained with the binary cross-entropy loss:

$$\mathcal{L}_{\text{probe}}(\phi) = \mathbb{E}_{x,y}\big[-y \log \hat{y} - (1-y)\log(1-\hat{y})\big]. \tag{1}$$

The probe learns a mapping from intermediate hidden states to the correctness label Y and therefore serves as a routing estimator. Compared with sequence-level signals such as input difficulty or output consistency, probes can capture richer semantic information encoded in hidden states (Chuang et al., 2025), which leads to effective routing performance.

AUC	MMMU	AOKVQA	ScienceQA
Text-only	**81.51**	**87.50**	**95.36**
multimodal	76.51	82.03	83.95

Table 1: Routing AUC of probes on multimodal datasets. Probes learned on text-only samples consistently outperform probes learned on multimodal inputs.

3.3 Probes Failed in MLLMs.

Although hidden-state probes are effective in text-only LLM routing, their performance degrades substantially in MLLMs. To examine this effect, we conduct two groups of experiments.

First, we use ScienceQA (Saikh et al., 2022), which naturally contains two subsets: a text-only subset and a subset with images. We train and evaluate probes separately on these two subsets. The results show that probe performance, measured by AUC, is consistently lower when images are present. Second, to reduce potential bias caused by differences in the underlying samples, we conduct an additional comparison on A-OKVQA (Schwenk et al., 2022) and MMMU (Yue et al., 2024), both of which originally contain only image-text paired samples. For each dataset, we construct a text-only version by prompting GPT to generate descriptions of the images. We then train and evaluate probes on both the original multimodal version and the caption-based text-only version. The results again show that probes perform better on the text-only versions, indicating that the presence of images weakens correctness signals in hidden states and leads to lower routing accuracy.

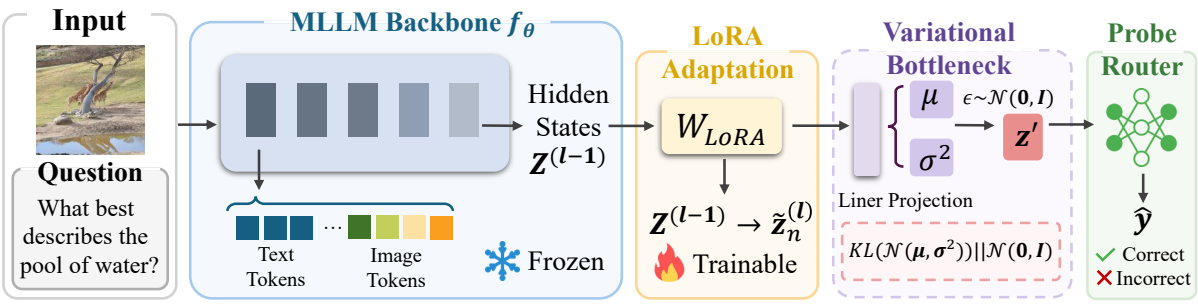


Figure 1: **ReLope pipeline.** Given an input sample $(\mathbf{x}, y)$, the MLLM f_θ produces hidden states $\mathbf{Z}^{(l-1)}$ at layer $l-1$. A lightweight LoRA adapter θ_{LoRA} is inserted into the model to obtain adapted hidden states $\tilde{\mathbf{Z}}^{(l)}$. The adapted last-token representation $\mathbf{z} = \tilde{\mathbf{z}}_n^{(l)}$ is then used as the routing feature. It is fed into the probe g_ϕ to predict correctness $\hat{y}$ for routing. In parallel, two linear heads map $\mathbf{z}$ to $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$, which parameterize a Gaussian posterior for a variational bottleneck. The training objective combines cross-entropy loss with a KL regularizer, encouraging the learned representation to preserve correctness-related information while filtering out task-irrelevant variation.

4 Method

4.1 Aggregation Hidden States for Probing

The empirical results in Sec. 3.3 show that the last token hidden state $\mathbf{z}_n^{(l)}$, which is the standard input to probes, provides a weak routing signal in MLLMs. One possible reason is that information relevant to answer correctness is not concentrated in the final token alone. Instead, such information can be distributed across multiple token representations, especially when textual and visual information are jointly encoded.

To address this issue, we construct the probe feature from token-level hidden states in the preceding layer. Let $\mathbf{Z}^{(l-1)} = \{\mathbf{z}_i^{(l-1)}\}_{i=1}^n$ denote the hidden states at layer $l-1$. In a transformer, the last token hidden state at layer l is produced by aggregating information from all tokens' hidden states in layer $l-1$ through the attention module. However, this aggregation is optimized for answer generation by the MLLM, rather than for routing. As a result, the resulting representation may emphasize information useful for producing the next token while weakening signals that are more informative for predicting answer correctness. In MLLMs, this issue can be more severe because irrelevant or redundant visual content may also be mixed into the final-token representation. We therefore propose to replace this implicit aggregation with a learned aggregation tailored to routing.

Attention Probe. We define the probe input as

$$\hat{\mathbf{z}} = \sum_{i=1}^{n} \beta_i \mathbf{z}_i^{(l-1)},$$

where β_i denotes the contribution weight of token i. The weights are computed with a lightweight attention mechanism:

$$s_i = \frac{\mathbf{q}^\top \mathbf{z}_i^{(l-1)}}{\sqrt{d}}, \quad \beta_i = \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)},$$

where $\mathbf{q} \in \mathbb{R}^d$ is a learnable routing query vector and d is the hidden dimension. The aggregated representation $\hat{\mathbf{z}}$ is then used as the input to the probe for routing.

This design does not modify the backbone MLLM and only changes how the probe reads intermediate representations. In this way, the probe can focus on routing-relevant cues while reducing the influence of irrelevant visual information that may be carried into the hidden state $\mathbf{z}_n^{(l)}$.

The results in Table 2 show that the Attention Probe consistently improves routing AUC over the standard last token probe across most datasets and backbone models. These results suggest that correctness-related signals in MLLMs are not well captured by the final-token representation alone, and that a learned token-level aggregation provides a stronger feature for routing.

4.2 ReLope

The Attention Probe reweights existing features but does not alter the representation space itself. We next consider a more expressive approach that directly adapts the hidden state for routing. To this end, we introduce **ReLope** (KL-Regularized LoRA Probe), which combines LoRA adaptation with a variational bottleneck objective. The LoRA adapters provide an efficient way to reshape the hidden representation without full fine-tuning, while

Algorithm 1: Attention Probe and ReLope for Routing

Input: Small MLLM f_θ, probe g_ϕ, optional LoRA at layer l, linear heads $(\mathbf{W}_\mu, \mathbf{b}_\mu)$ and $(\mathbf{W}_{\log\sigma^2}, \mathbf{b}_{\log\sigma^2})$, training set $\mathcal{D}$, learning rate η, epochs T, KL weight schedule $\beta(t)$

Output: Trained parameters $(\theta_{\text{LoRA}}, \mathbf{W}_\mu, \mathbf{b}_\mu, \mathbf{W}_{\log\sigma^2}, \mathbf{b}_{\log\sigma^2}, \phi)$

for $t = 1$ **to** T **do**
  for *mini-batch* $\{(\mathbf{x}_b, y_b)\}_{b=1}^{B} \subset \mathcal{D}$ **do**
    `// Attention Probe (Sec. 4.1)`
    $\mathbf{Z}_b^{(l-1)} \leftarrow f_\theta(\mathbf{x}_b)$;
    $\hat{\mathbf{z}}_b \leftarrow \sum_i \beta_i \mathbf{z}_{b,i}^{(l-1)}$;
    $\hat{y}_b^{\text{attn}} \leftarrow \sigma\big(g_\phi(\hat{\mathbf{z}}_b)\big)$;
    $\mathcal{L}_{\text{Attn}} \leftarrow \frac{1}{B}\sum_b \text{BCE}(y_b, \hat{y}_b^{\text{attn}})$;
    Update (ϕ) to minimize $\mathcal{L}_{\text{Attn}}$;
    `// ReLope (Sec. 4.2)`
    $\tilde{\mathbf{Z}}_b^{(l)} \leftarrow f_\theta(\mathbf{x}_b)$ with LoRA;
    $\mathbf{z}_b \leftarrow \tilde{\mathbf{z}}_{b,n}^{(l)}$;
    $\boldsymbol{\mu}_b \leftarrow \mathbf{W}_\mu \mathbf{z}_b + \mathbf{b}_\mu$;
    $\log \boldsymbol{\sigma}_b^2 \leftarrow \mathbf{W}_{\log\sigma^2}\mathbf{z}_b + \mathbf{b}_{\log\sigma^2}$;
    $\hat{y}_b^{\text{kl}} \leftarrow \sigma\big(g_\phi(\mathbf{z}_b)\big)$;
    $\mathcal{L}_{\text{CE}} \leftarrow \frac{1}{B}\sum_b \text{BCE}(y_b, \hat{y}_b^{\text{kl}})$;
    $\mathcal{L}_{\text{KL}} \leftarrow D_{KL}(q_\psi \| \mathcal{N}(0, I))$;
    $\mathcal{L}_{\text{ReLope}} \leftarrow \mathcal{L}_{\text{CE}} + \beta(t)\mathcal{L}_{\text{KL}}$;
    Update $(\theta_{\text{LoRA}}, \mathbf{W}_\mu, \mathbf{b}_\mu, \mathbf{W}_{\log\sigma^2}, \mathbf{b}_{\log\sigma^2}, \phi)$ to minimize $\mathcal{L}_{\text{ReLope}}$;

the bottleneck encourages the adapted feature to retain information related to answer correctness and discard irrelevant variation.

Model Structure. We insert LoRA adapters into the selected transformer layer l and keep the original backbone parameters frozen. The LoRA adapters take the hidden states $\mathbf{Z}^{(l-1)} \in \mathbb{R}^{n\times d}$ at transformer layer $l-1$ to produce adapted hidden states $\tilde{\mathbf{Z}}^{(l)}$, from which we can get the last-token representation

$$\mathbf{z} = \tilde{\mathbf{z}}_n^{(l)}$$

as the routing feature. Instead of sending $\mathbf{z}$ directly to the probe classifier, we first map it to a stochastic bottleneck. Two linear projection heads produce the parameters of a diagonal Gaussian:

$$\boldsymbol{\mu} = \mathbf{W}_\mu \mathbf{z} + \mathbf{b}_\mu, \quad \log \boldsymbol{\sigma}^2 = \mathbf{W}_{\log\sigma^2}\mathbf{z} + \mathbf{b}_{\log\sigma^2}.$$

These parameters define a variational posterior

$$q_\psi(\mathbf{z}' \mid \mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)),$$

from which the compressed representation is sampled using the reparameterization trick:

$$\mathbf{z}' = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

The probe prediction is then computed from the compressed feature:

$$\hat{y} = \sigma(g_\phi(\mathbf{z}')).$$

At inference time, we use the posterior mean $\boldsymbol{\mu}$ in place of sampling.

Training Objective. ReLope jointly optimizes probe prediction and representation compression:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x},y)}\big[-y\log\hat{y} - (1-y)\log(1-\hat{y})\big] + \beta D_{\text{KL}}\left(\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\right). \tag{2}$$

The first term is the binary cross-entropy loss for routing prediction. The second term regularizes the bottleneck representation by encouraging the posterior to stay close to a standard Gaussian prior. Following the information bottleneck view (Tishby and Zaslavsky, 2015; Alemi et al., 2016), this regularization encourages the representation to preserve information that is useful for predicting correctness while discarding irrelevant details in the hidden states. The coefficient β controls the strength of the compression.

Unlike existing routing methods, which make routing decisions from fixed model inputs or outputs, ReLope learns routing-oriented features by adapting the hidden states themselves. This design allows ReLope to make decisions based on features that are more informative for the answer's correctness in MLLMs and more robust to irrelevant multimodal variation.

5 Experiment

5.1 Datasets

We evaluate ReLope routing performance on five widely used multimodal reasoning benchmarks. **MMMU** (Yue et al., 2024) is designed to evaluate advanced multimodal reasoning across multiple academic disciplines and contains thousands of expert-level questions spanning more than 30 subjects. **AOKVQA** (Schwenk et al., 2022) contains over 25,000 questions and requires both external

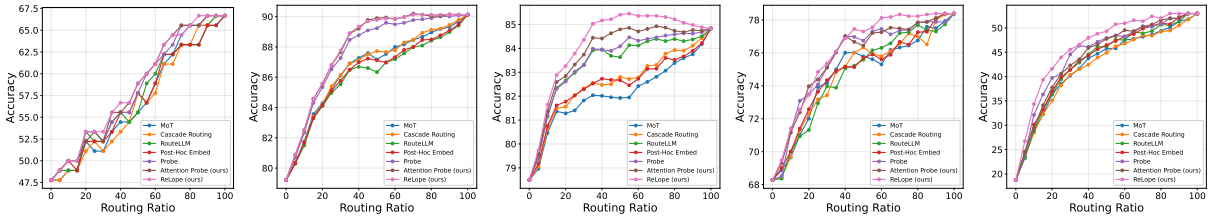
Method	MMMU	AOKVQA	ScienceQA	ChartQA	MathVision	Avg.
Qwen2.5-VL-7B-Instruct						
MoT	67.52	72.35	72.08	75.38	82.59	73.18
Cascade Routing	68.32	73.15	76.31	69.58	78.25	73.12
RouteLLM	71.33	68.51	82.83	74.89	87.53	77.02
Post-Hoc Embed	70.30	69.92	75.45	72.75	88.14	75.31
Probe	76.51	82.03	83.95	82.51	90.28	83.86
Attention Probe (ours)	77.10	84.34	86.43	85.43	88.92	84.84
ReLope (ours)	**80.89**	**86.17**	**91.79**	**86.51**	**95.68**	**88.21**
Gemma3-12B						
MoT	62.37	68.05	72.84	70.16	65.39	67.76
Cascade Routing	62.52	64.91	75.58	73.29	68.10	68.88
RouteLLM	64.83	66.10	78.92	74.65	70.45	70.59
Post-Hoc Embed	65.68	71.92	76.19	73.42	69.07	71.26
Probe	68.24	75.13	80.47	76.88	72.26	74.60
Attention Probe (ours)	70.39	74.58	84.93	81.32	71.19	76.48
ReLope (ours)	**76.23**	**85.73**	**88.73**	**87.95**	**81.77**	**84.88**
Phi-4-Multimodal-Instruct						
MoT	63.91	68.42	74.03	67.17	73.28	69.36
Cascade Routing	63.79	70.04	75.56	62.93	76.84	69.83
RouteLLM	66.12	70.81	84.13	70.46	78.55	74.41
Post-Hoc Embed	67.28	72.66	77.84	70.58	77.12	73.10
Probe	69.97	76.23	82.08	73.76	81.03	76.61
Attention Probe (ours)	72.68	80.27	87.24	78.63	83.47	80.46
ReLope (ours)	**78.75**	**85.90**	**92.36**	**81.47**	**91.16**	**85.93**

Table 2: AUC (%) comparison of routing methods on five multimodal benchmarks: MMMU, AOKVQA, ScienceQA, ChartQA, and MathVision. Results are reported for three backbone MLLMs: Qwen2.5-VL-7B-Instruct, Gemma3-12B, and Phi-4-Multimodal-Instruct. Higher values indicate better routing quality.

world knowledge and image understanding ability to answer correctly. **ScienceQA** (Saikh et al., 2022) is a multimodal science question answering benchmark consisting of over 21,000 middle school science problems. **ChartQA** (Masry et al., 2022) is a benchmark for question answering over charts and tables, requiring models to interpret structured visual data and perform logical reasoning with it. **MathVision** (Wang et al., 2024) contains visually grounded mathematics problems, requiring models to interpret diagrams and symbolic expressions in order to solve complex mathematical questions.

5.2 Baselines

We compare ReLope with several representative routing strategies. **MoT** (Yue et al., 2023) performs routing based on multiple reasoning traces generated by the model. The final routing decision is obtained by aggregating signals from different reasoning paths, which provides a more stable estimate of task difficulty. **Cascade Routing** (Chen et al., 2023) adopts a staged routing framework in which inputs are first processed by a lightweight model and are escalated to stronger models only when necessary. We implement Strategy 3 from the original paper, which determines routing decisions based on model confidence signals. **RouteLLM** (Ong et al., 2024) learns a routing function that ranks candidate models according to expected performance and cost. Following the original work, we use the SW Ranking strategy to estimate routing preferences between models. **Post-Hoc Embed** (Gupta et al., 2024) extracts the output embedding of the small model after generation. A lightweight classifier is trained on these embeddings to estimate whether the small model can correctly solve the input problem. **Probe** (Chuang et al., 2025) trains a supervised probe on intermediate hidden states



(a) MMMU (b) AOKVQA (c) ScienceQA (d) ChartQA (e) MathVision

Figure 2: Overall system accuracy as a function of routing ratio. The routing ratio denotes the percentage of queries routed to the large model in a hybrid MLLM system. Higher curves indicate better routing strategies that achieve stronger trade-offs bewteen accuracy and cost by sending only the most difficult queries to the large model.

of the model to estimate correctness probability, which serves as the routing signal.

5.3 Implementation Details

All experiments are implemented in PyTorch using the transformers and peft libraries. For probe-based methods (Probe, Post-Hoc Embed, Attention Probe, and ReLope), a five-layer MLP probe is used. Training uses the AdamW optimizer with learning rate 1×10^{-4}. Hidden states are extracted from an intermediate transformer layer l of the small model f_θ. Layer $l = 24$ is used for Qwen2.5-VL-7B-Instruct and Phi-4-Multimodal-Instruct (32 layers), and layer $l = 40$ for Gemma3-12B (48 layers). For ReLope, a LoRA adapter with rank $r = 128$ and scaling factor $\alpha = 256$ is inserted into the selected transformer layer. Two linear heads predict $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}^2$ from the adapted feature $\mathbf{z}$. The loss combines binary cross-entropy and KL divergence with $\beta = 1.0$. All experiments run on NVIDIA A100 GPUs with 80 GB memory.

5.4 Comparison with Baselines

AUC performance comparison. Table 2 reports routing performance on five datasets and three backbone MLLMs: Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Gemma3-12B (Team et al., 2025), and Phi-4-Multimodal-Instruct (Abouelenin et al., 2025). The Attention Probe and ReLope consistently achieve the best results across all datasets and backbone models. On Qwen2.5-VL-7B-Instruct, ReLope achieves an average AUC of 88.97, improving over Attention Probe by 4.13. Similar improvements appear on Gemma3-12B and Phi-4-Multimodal-Instruct, where the average AUC achieves 84.88 and 85.93, respectively. These results indicate that hidden state reconstruction improves the routing performance.

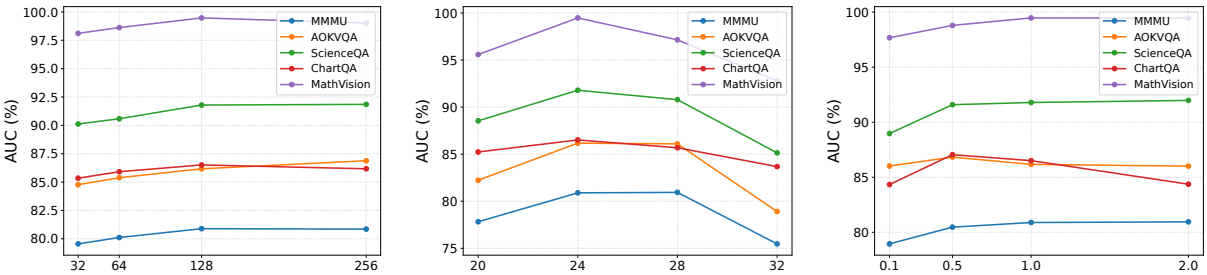
Routing performance in the hybrid MLLM system. Following (Chuang et al., 2025), we evaluate the trade-off between accuracy and cost in a hybrid MLLM system composed of Qwen2.5-VL-7B-Instruct and a stronger large model. We use GPT-4.1 for most datasets and GPT-5.1 (OpenAI, 2026) for MathVision. Given a routing ratio of $h\%$, each method ranks queries by its routing score and sends the top-$h\%$ hardest queries to the large model, while the remaining queries are answered by the small model.

Figure 2 reports overall system accuracy as the routing ratio varies. When the ratio is 0, all queries are handled by the small model; when it is 100, all queries are routed to the large model. ReLope consistently traces the strongest tradeoff across all datasets. Its advantage is most evident at low routing ratios, where it gains higher accuracy under the same budget. This shows that ReLope is better at identifying queries that truly require escalation, leading to more efficient hybrid MLLM routing.

5.5 Ablations and Discussion

Robustness to visual perturbations. To further examine whether ReLope suppresses task-irrelevant visual noise, we evaluate routing robustness under image perturbations. For each test sample in AOKVQA, we apply several typical perturbations to the input image, including Gaussian noise, JPEG compression, and Gaussian blur. The text prompt and ground-truth label remain unchanged. To summarize robustness across perturbation types, we compute the average AUC drop relative to clean inputs: $\Delta\text{AUC} = \text{AUC}_{clean} - \frac{1}{K}\sum_{k=1}^{K} \text{AUC}_{perturb}^{(k)}$, where K denotes the number of perturbation types. A smaller ΔAUC indicates stronger robustness.

Table 3 shows that ReLope consistently exhibits the strongest robustness across perturbations. While all routing methods experience performance degradation when visual noise is introduced, the drop is substantially smaller for ReLope. The result



(a) LoRA rank r (b) Transformer layer l (c) VIB coefficient β

Figure 3: Ablation studies of ReLope. (a) Effect of the LoRA rank r. (b) Effect of the transformer layer l used for probe placement. (c) Effect of the VIB coefficient β. Results are reported as AUC on five multimodal benchmarks.

Method	Clean	Noise	JPEG	Blur	ΔAUC↓
Probe	82.03	77.31	76.84	76.05	5.30
Attention Probe (ours)	84.24	82.11	81.65	80.90	2.69
ReLope (ours)	**86.17**	**86.05**	**84.21**	**83.92**	**1.44**

Table 3: Routing robustness under visual perturbations. We report routing AUC under image perturbations. ΔAUC denotes the average AUC drop across perturbations relative to clean inputs.

Method	MMMU	AOKVQA	ScienceQA	ChartQA	MathVision
MoT	67.52	72.35	72.08	75.38	**82.59**
Cascade Routing	67.15	72.94	75.33	65.47	63.28
RouteLLM	69.28	65.57	75.98	64.38	61.16
Post-Hoc Embed	67.55	66.18	69.81	65.75	64.39
Probe	71.78	81.92	80.57	73.05	66.35
Attention Probe (ours)	73.22	81.87	81.14	73.98	66.42
ReLope Probe (ours)	**76.66**	**83.68**	**81.58**	**75.57**	66.83

Table 4: OOD evaluation of different probes. ReLope achieves the best generalization and suffers the smallest performance drop under domain shift.

suggests that the LoRA adaptation with KL regularization encourages the router to learn representations related to routing filter out task-irrelevant visual noise. As a result, ReLope produces more stable routing decisions.

Generalization on OOD data. We evaluate routing generalization under domain shift. For each target dataset, the probe is trained on the remaining four datasets and evaluated on the chosen dataset. Table 4 shows that ReLope achieves the best performance on four out of five benchmarks, consistently outperforming existing probe-based methods, which indicates stronger cross-dataset generalization. However, the gain on MathVision is smaller. This is expected because MathVision focuses on mathematical reasoning, which differs substantially from the general QA tasks represented by the other datasets. The resulting domain gap makes transferring routing signals learned from general QA datasets more challenging. Nevertheless, ReLope still achieves the best performance among probe-based methods on MathVision, suggesting that the learned routing-aware representations remain more robust under substantial domain shift.

Parameter ablation. Figure 3 shows the impact of three key parameters: the LoRA rank r, the probe layer l, and the VIB coefficient β. Increasing the LoRA rank improves routing AUC up to $r = 128$, while further increases provide only marginal gains. This suggests that a moderate rank is sufficient to adapt the hidden representation for routing. For probe placement, intermediate and later transformer layers provide stronger routing signals than early layers, indicating that these layers contain more informative features for correctness prediction. For the VIB coefficient, moderate values ($\beta \in [0.5, 1.0]$) achieve the best performance, suggesting a suitable balance between information compression and predictive accuracy.

6 Conclusion

This work studies routing in MLLMs and shows that probe-based methods, which are effective in text-only LLMs, degrade substantially in the multimodal setting. To address this issue, two methods are proposed. The Attention Probe aggregates token-level hidden states to enrich the probe input, while ReLope introduces a LoRA adapter with a KL regularizer to learn routing-aware representations. Extensive experiments across multiple benchmarks and backbones show consistent improvements over strong baselines. In hybrid systems, our proposed methods achieve a better accuracy–cost trade-off, especially under low routing budgets. These results highlight the importance of learning a routing-aware probe in MLLMs. More broadly, the findings suggest that improving the structure of hidden states is a promising direction for reliable routing in multimodal systems.

7 Limitations

This work has several limitations. First, although the experiments cover multiple benchmarks and backbone MLLMs, the evaluation is still limited to a small set of tasks and model families. It remains unclear how well the proposed methods generalize to other multimodal settings and broader deployment scenarios. Second, ReLope relies on supervised correctness labels and introduces additional tuning overhead, which may limit scalability in new domains. Finally, while we show that standard hidden-state probes are much less effective in MLLMs than in text-only LLMs, the underlying reason is still not fully understood. Our results suggest that multimodal hidden states contain weaker routing signals, but a deeper analysis of why probe methods fail in MLLMs remains necessary.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Dingqi Yang, Hailong Sun, and Philip S Yu. 2025. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036*.

Yu-Neng Chuang, Leisheng Yu, Guanchu Wang, Lizhe Zhang, Zirui Liu, Xuanting Cai, Yang Sui, Vladimir Braverman, and Xia Hu. 2025. Confident or seek stronger: Exploring uncertainty-based on-device llm routing from benchmarking to generalization. *arXiv preprint arXiv:2502.04428*.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Language model cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, and 1 others. 2025. Universal model routing for efficient llm inference. *arXiv preprint arXiv:2502.08773*.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd international conference on computer, artificial intelligence and control engineering*, pages 405–409.

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. Routing to the expert: Efficient reward-guided ensemble of large language models. *arXiv preprint arXiv:2311.08692*.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. *arXiv preprint arXiv:2406.13415*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. 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.

Alireza Mohammadshahi, Arshad Rafiq Shaikh, and Majid Yazdani. 2024. Routoo: Learning to route to large language models effectively. *arXiv preprint arXiv:2401.13979*.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*.

OpenAI. 2026. Chatgpt. https://chat.openai.com/.

Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. 2025. Adaptive llm routing under budget constraints. *arXiv preprint arXiv:2508.21141*.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly

articles. *International Journal on Digital Libraries*, 23(3):289–301.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Soham Shah and Kumar Shridhar. 2025. Select-then-route: Taxonomy guided routing for llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 425–441.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. 2023. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. 2023. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning. *arXiv preprint arXiv:2310.03094*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Haozhen Zhang, Tao Feng, and Jiaxuan You. 2025. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zesen Zhao, Shuowei Jin, and Z Morley Mao. 2024. Eagle: Efficient training-free router for multi-llm inference. *arXiv preprint arXiv:2409.15518*.