CoGR-MoE: Concept-Guided Expert Routing with Consistent Selection and Flexible Reasoning for Visual Question Answering

Anonymous ACL submission

Abstract

Visual Question Answering (VQA) requires models to identify the correct answer options based on both visual and textual evidence. Recent Mixture-of-Experts (MoE) methods improve option reasoning by grouping similar concepts or routing based on examples. However, unstable routing can lead to inconsistent expert selection in the same question type, while overly stable routing may reduce flexibility. To address this, we propose Concept-Guided Routing framework (CoGR-MoE), which incorporates semantics of the answer options to guide expert selection in the training phase. Next, option features are used to reweight the selected experts, producing discriminative representations for each candidate option. These option-level representations are further used for option comparison and optimized via contrastive learning. The experimental results indicate that CoGR-MoE delivers strong performance across multiple VQA tasks, demonstrating the effectiveness of our approach.

1 Introduction

VQA is a multimodal reasoning task where a model should identify the correct answer among several candidates by grounding its decision in visual and textual evidence(Kim et al., 2025). Flexible multimodal reasoning is essential for real-world applications including visual assistants, robotics and accessibility tools(Borisova et al., 2025). However, such fine-grained reasoning remains challenging due to the diversity and complexity of visual–language representations. Recent advances incorporate MoE architectures into vision–language models to allow different experts to specialize in distinct visual or textual patterns(Le et al., 2025; Nakamura et al., 2025).

The effectiveness of MoE models heavily depends on routing decisions, which can become inconsistent for inputs that require the same answer but are expressed differently.(Olson et al., 2025). Recent work has begun to address this routing instability by improving routing mechanisms. Some methods like MoKE(Cheng et al., 2025b) assign similar concepts to the same knowledge expert, while other approaches(Li et al., 2025a) like ERMoE(Cheng et al., 2025a) routes inputs based on their similarity to the experts' representations. Although these methods maintain routing stability across questions of the same type, such stability also solidifies the activated expert set, thereby weakening flexibility in option-level discrimination(Duanmu et al., 2025; Li et al., 2025c). As a result, the model struggles to capture the distinct evidence required by each option, leading to errors in comparative reasoning.

Expert roles in MoE models emerge through repeated activation under consistent semantic conditions, requiring cues that remain stable across different representations. Once the Top-K expert set is static, any remaining discriminative capacity come from how expert outputs are utilized. Accordingly, dynamically adjusting the relative contributions of experts can help improve the model's ability to distinguish among candidate answers.

Guided by the above considerations, we propose CoGR-MoE, as illustrated in Figure 1, to mitigate the imbalance between routing instability and expert utilization rigidity. First, positive and negative feature descriptions are generated by a Large Language Model (LLM) for each option. The features of the correct answer are injected into the MoE gating logits as a conceptual direction. After the Top-K experts are selected, feature descriptions from options are used to reweight expert contributions, enabling distinct representations for each option. These option representations are optimized with contrastive learning in the training phase. The experiments show that CoGR-MoE outperforms several strong baselines, indicating the benefit of balancing routing stability with flexible expert utilization.

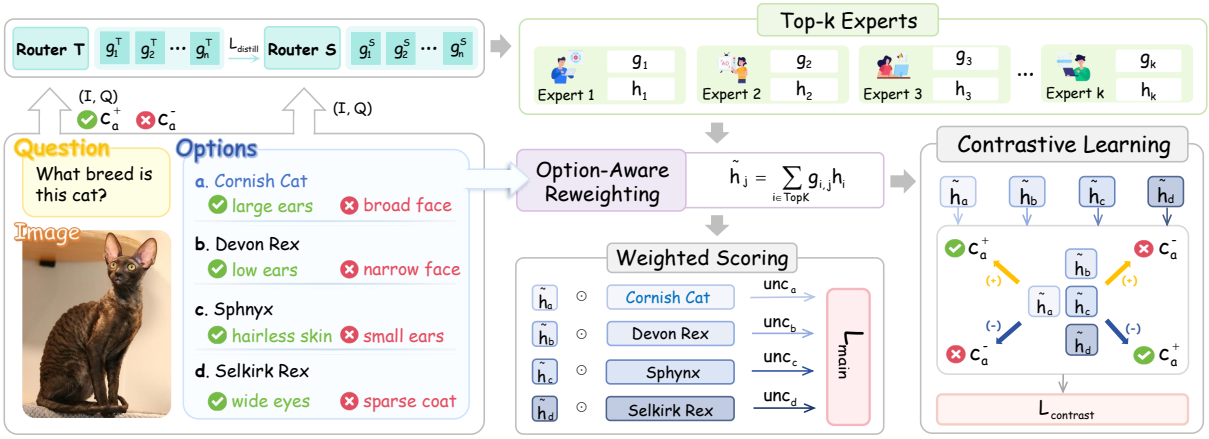

Figure 1: The CoGR-MoE framework incorporates features of the correct answer c_a^+ and c_a^- into the router. This process also produces teacher gating weights g^T from Router T, which are distilled into student gating weights g^S of Router S. It then applies semantics of option j to flexibly modulate expert contributions $g_{i,j}$, where i indexes the selected experts. The expert outputs h_i are aggregated using gating weights $g_{i,j}$, yielding features $\tilde{h}_j$ for each option. Contrastive learning is employed with positive and negative option pairs c_a^+ and c_a^- to identify correct options.

Our contributions are summarized as follows:

- We propose CoGR-MoE, a concept-guided MoE framework that stabilizes expert selection by injecting answer-relevant features into the routing process.
- CoGR-MoE also adjusts expert contributions by option content, enhancing option-level discrimination while preserving routing consistency.
- Extensive experiments on datasets validate the efficacy of CoGR-MoE, achieving up to a 4.9% absolute gain in accuracy, with more interpretable expert utilization.

2 Related Works

2.1 Embedding-Driven Routing

Embedding-driven routing computes expert scores from an input's hidden representation, and typically relies on representation-driven mechanisms to stabilize outputs(Li et al., 2025b). MoME (Shen et al., 2024a) partitions experts into modality-specific and shared groups, making inputs from the same modality more likely to activate the same experts. In parallel, MH-MoE (Huang et al., 2024)routes multiple sub-representations of a token to different experts and aggregates their outputs, yielding stable expert combinations despite routing variability. In MMOE(Shen et al., 2024b), similar inputs mapped to nearby points in the joint embedding space are assigned similar expert weights. However, a limitation is that routing is based solely on joint embeddings, without distinguishing modality interactions or semantic roles, which may reduce representation alignment and interpretability in expert selection.

2.2 Interaction-Aware Routing

To address the weakened correspondence between routing decisions and cross-modal structure, interaction-driven gating leverages task-specific and cross-modal interaction features to guide expert selection(Cai et al., 2025)(Han et al., 2025). I^{2}MoE(Xin et al., 2025) select expert on interaction types, such that inputs exhibiting the same modality interaction patterns tend to activate the same group of experts. Similarly, MOEMOE (Verma et al., 2025) guide expert selection with query-aware cross-modal attention, thereby inputs with comparable modality relevance patterns tend to activate similar experts. RoE-LLaVA(Wu et al., 2025) applyies a contrastive objective that pulls semantically related samples closer in the routing embedding space, encouraging them to share relevant expert routes. These methods mitigate routing shifts caused by interactions, but do not address routing inconsistency when semantically similar inputs are mapped to different embeddings(Mu and Lin, 2025).

2.3 Semantic-Informed Routing

To further mitigate inconsistent expert assignment under representation variation, semantic-informed routing aligns expert selection across semantically related samples. MoKE(Cheng et al., 2025b)

uses projection-based routing to consistently assign knowledge edits involving similar concepts to the same group of domain-specific experts, while R2-T2 (Li et al., 2025c) adjusts routing by comparing new samples with similar past samples. ProMoE (Wei et al., 2025) introduces a two-step routing mechanism that first separates tokens by functional roles and then assigns them to experts based on similarity to learnable vectors. These methods are built on the mechanism that similar inputs should be routed to the same or similar sets of experts (Li et al., 2025d). As a result, the model tends to reuse the same experts even when the question undergoes subtle variations, which can misjudge options.

3 Methodology

3.1 Framework Overview

Existing MoE routing methods(Cheng et al., 2025b; Li et al., 2025d) achieve more consistent expert selection within the same problem type through semantic grouping. However, this stability can cause routing to become static, repeatedly selecting the similar experts even when the question varies, leading to incorrect answer selection. To address this limitation, we propose CoGR-MoE, which incorporates features of answers option into the gating mechanism to align expert selection with the correct answer. Meanwhile, CoGR-MoE reweights the selected experts using features of options, improving discrimination and reducing incorrect answer selection.

In CoGR-MoE, the LLM generates positive and negative visual cues for each option, indicating which attributes must or must not appear for the choice to be correct. The correct answer's cues are then used to inject into the gating logits to guide the selection of Top-K expert set, as illustrated in Figure 2. Then each option uses its own cues to modulate the expert weights, reweighting the selected experts to form an option-specific representation. Finally, representations are used for answer prediction and optimized with a cue-based contrastive loss in the training phase.

3.2 Semantic Cues Generation

To provide stable evidence descriptions for routing, an LLM generates positive and negative semantic cues for each answer choice. Positive cues describe what must appear in the image for an option to be correct, while negative cues specify what must not. To quantify how well an option's cues align with the image, we introduce an agreement score Agr. Agr combines the image's similarity to positive cues with an inversely weighted similarity to negative cues. In parallel, paraphrased variants of the cues are generated via synonym substitutions, and their cue-image similarities are used to compute a variance score Var_j. Finally, the uncertainty of option j is defined as a normalized ratio of Var to its Agr:

$$unc = \frac{Var}{1 + Agr}. \tag{1}$$

When Var is high, the uncertainty increases because the cues exhibit unstable alignment with the image. Conversely, when Agr is high, the uncertainty decreases since the answer is strongly supported by the image semantics. For the answer option, if the estimated uncertainty unc exceeds a predefined threshold, semantic cues are regenerated to maintain reliable cue–image alignment.

3.3 Concept-Guided Expert Routing

Cues are more stable than specific phrasing because they capture decision-relevant attributes rather than textual variation. As a result, to inject stable guidance into the routing process, we construct a semantic direction s_a from the cues of the correct option a. It is obtained by taking the difference between the embeddings of positive cues c_a^+ and negative cues c_a^- to support the correct features while suppressing conflicting attributes.

$$s_a = W_{\text{sem}}(c_a^+ - c_a^-). \tag{2}$$

The base logits z_{base} are the expert scores produced by the MoE router using only the representation of image–question. The semantic direction s_a is added to z_{base} with a fixed routing strength $\lambda_a = 0.5$ to obtain the concept-guided gating distribution:

$$g^T = \text{softmax}(z_{\text{base}} + \lambda_a\, s_a)\,. \tag{3}$$

where g^T denotes the gating distribution produced by the teacher router over the Top-K experts. The teacher and student routers are denoted as Router T and Router S, respectively. The Top-K experts selected by g^T are encouraged to align with the semantic direction implied by the cues of the correct option. To enable cue-free inference, Router S with the same architecture operates without s_a. Its gating distribution is computed as

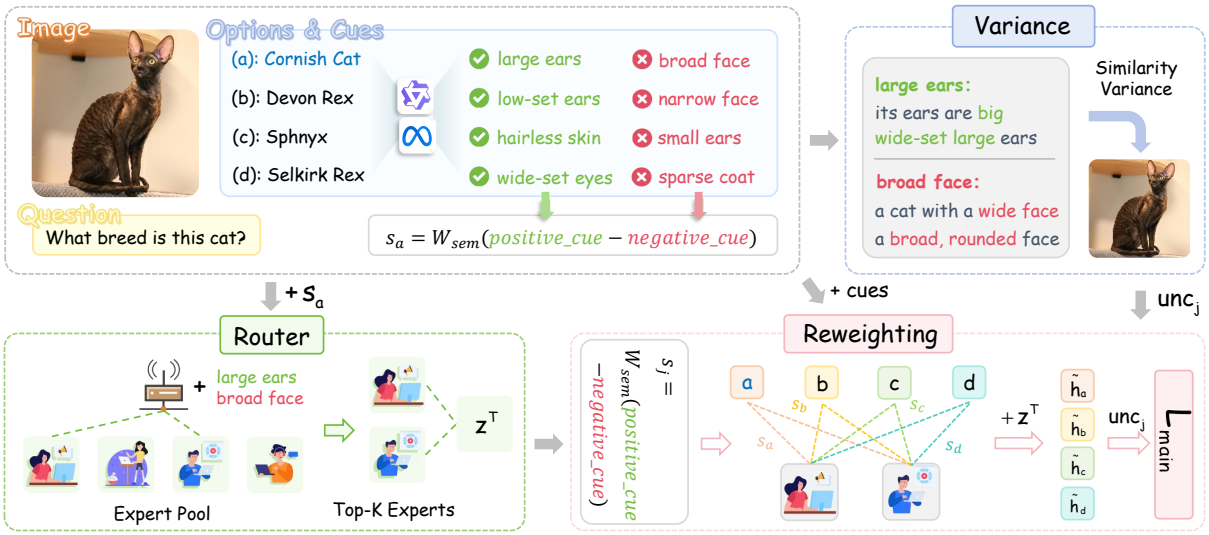

Figure 2: Features of answer option s_a are injected into the router to guide consistent Top-K expert selection, producing initial routing logits z^T. The uncertainty estimate unc_j measures the reliability of the s_j, which means the features of each option. Together, s_j adjust expert weights to form the option representation $\tilde{h}_j$. L_{main} applies unc_j-weighted cross-entropy to option representations $\tilde{h}_j$.

$$g^S = \mathrm{softmax}(z_{\mathrm{base}}), \tag{4}$$

where g^S denotes the gating distribution of Router S. This process does not require any cue generation. During training, a KL-divergence loss is introduced to distill the routing distribution of Router T into the Router S:

$$L_{distill} = \mathrm{KL}\big(g^T \,\|\, g^S\big)\,. \tag{5}$$

This distillation encourages the Router S to approximate the behavior of Router T, enabling it to select experts along the answer-relevant direction at inference without requiring any cues.

3.4 Option-Aware Expert Reweighting

After the shared Top-K experts are selected, an option-specific allocation step is required to maintain discriminability among answer choices. To achieve this, features of each option j are derived from its positive and negative cues to compute s_j, which is constructed in the same manner as the routing guidance in Eq. (2). $s_{i,j}$ is injected into the logits of the Top-K experts i, adjusting their weights so that experts more relevant to the current option receive higher importance. z_i^T denotes the i-th component of the Router T routing logits z^T before option-aware modulation.

$$g_{i,j} = \mathrm{softmax}_{i\in\mathrm{TopK}}\big(z_i^T + \lambda_o\, s_{i,j}\big)\,, \tag{6}$$

where $g_{i,j}$ is defined as the i-th element of g_j, representing the option-specific gating distribution over the Top-K experts i. λ_o controls the strength of reweighting and is fixed to 0.5. Each option then forms its own aggregated representation $\tilde{h}_j$ by weighting the shared expert outputs according to its gating distribution $g_{i,j}$:

$$\tilde{h}_j = \sum_{i\in\mathrm{TopK}} g_{i,j}\, h_i, \tag{7}$$

where i denotes the expert index within the Top-K routed experts, and h_i is the output of the i-th expert. This mechanism allows all options to share one expert set while still producing distinct representations. If each option were allowed to reweight all experts, the routing would drift and become unstable. Finally, the predictive score $score_j$ for option j is computed as the cosine similarity between its aggregated representation $\tilde{h}_j$ and the embedding of its option text.

3.5 Training and Inference

The L_{main} relies on $score_j$ for each option. However, because different options exhibit different levels of uncertainty, each option is weighted according to the confidence of its cues. y_j is the one-hot label indicating whether option j is the correct answer:

$$L_{main} = \sum_j \frac{1}{1+unc_j}\,\mathrm{CE}(score_j, y_j), \tag{8}$$

where CE denotes the cross-entropy loss. To further keep the model's internal representations aligned with decision-relevant semantics, we incorporate a cue-guided contrastive loss. The contrastive loss aligns the aggregated representation of the correct option with its positive cues, while encouraging the representations of incorrect options to be closer to the negative cues:

$$L_{contrast} = -\lambda_c \Big[\cos(\tilde{h}_{\text{correct}},\, c_a^{+}) - \cos(\tilde{h}_{\text{wrong}},\, c_a^{-}) \Big], \tag{9}$$

where $\tilde{h}_{\text{correct}}$ denotes the representation of the correct option, and $\tilde{h}_{\text{wrong}}$ is computed as a $score_j$-weighted average over incorrect options. λ_c is a constant value of 0.3 in all experiments. Finally, the complete training objective integrates all three components, including the $L_{distill}$ in Eq. (5) into a unified optimization target:

$$L_{total} = L_{main} + L_{contrast} + L_{distill}. \tag{10}$$

During the training phase, we construct a balanced multiple-choice supervision set by sampling 5,000 questions from each of IconQA(Lu et al., 2022b), A-OKVQA(Schwenk et al., 2022), and Visual7W(Zhu et al., 2016). The same training questions are used for all baselines to ensure a fair comparison.

During the inference phase, no cues are generated. CoGR-MoE relies solely on the Router S, which selects the Top-K experts based on the image–question pair. These experts are computed once and shared across all candidate answers. For each option j, the learned option-specific adjustment signal obtained from its text embedding is applied to allocate weights over the shared experts, producing an aggregated representation $\tilde{h}_j$. The model then computes a cosine similarity score between $\tilde{h}_j$ and the option text embedding, and the option with the highest score is chosen as the final answer.

4 Experiments

4.1 Experimental Setup

We evaluate CoGR-MOE on VMCBench(Zhang et al., 2025) and MRAG-Bench(Hu et al., 2024), two benchmarks for vision-language reasoning, and compare it with other MoE methods. VMCBench unifies twenty existing VQA datasets by converting open-ended questions into four-option multiple-choice, while MRAG-Bench is a vision-centric benchmark with images and annotated multiple-choice questions across several scenarios. Results are evaluated using task accuracy as the performance metric.

To further quantify the contribution of each component, we conduct an ablation study in which each variant disables one module while keeping remaining modules unchanged. Variants of CoGR-MoE include:

- **w/o $s_{\mathbf{a}}$:** removes the correct-answer semantic s_{a} and forms the routing anchor solely from image–question pair, eliminating explicit semantic grounding.
- **w/o s_j:** keeps the Top-K experts but removes the option-specific semantic term s_j, forcing all options to share the same routing-level expert weights, with option discrimination relying only on option text embeddings.
- **w/o unc_j:** removes uncertainty-aware weighting derived from cue consistency, and applies uniform weighting across all samples during training.
- **w/o $L_{contrast}$:** removes the contrastive loss that aligns expert representations with positive answer semantics while pushing them away from negative cues.
- **w/o $\mathcal{L}_{distill}$:** removes the distillation loss that transfers soft routing signals from the Router T to the Router S.
- **Prompt-only:** removes all cue-based training components, and uses the same LLM prompt only at inference phase to isolate the effect of cue-based training.

The influence of the MoE architecture on CoGR-MoE is examined by varying the total number of experts $n \in \{1, 2, 4, 8\}$ and the number of activated experts $K \in \{1, 2, 3, 4\}$ on LLaVA-1.5-7B. Each (n, K) configuration is evaluated using two metrics. Accuracy reflects practical performance, while the semantic–expert alignment score Sim measures how well expert routing follows the intended semantic direction.

$$\text{Sim} = \cos(h_{\text{Top}K},\, s_a)\,, \tag{11}$$

where s_a denotes the semantic direction of the correct answer, and $h_{\text{Top}K}$ represents the aggregated representation of the Top-K experts.

Table 1: Performance comparison of CoGR and other different MoE-based models on the MRAG-Bench. Results are reported for overall accuracy and across sub-categories including Perspective, Transformative, and Others. Detailed results for finer-grained sub-categories are provided in the Appendix A.

Methods	Overall	Perspective	Transformative	Others
MOE-LLaVA(Lin et al., 2024)				
MOE-LLaVA	53.29	55.54	50.78	51.67
MH-MoE (Huang et al., 2024)	61.71	69.78	51.70	40.83
Metis-HOME (Lan et al., 2025)	57.25	63.79	51.10	55.75
I^2MoE (Xin et al., 2025)	59.35	62.16	50.87	60.83
CL-MOE (Huai et al., 2025)	58.95	63.15	51.57	59.94
MoME (Shen et al., 2024a)	62.76	**69.30**	55.91	**64.06**
CoGR-MoE (Ours)	**63.25**	69.09	**58.55**	58.67
Qwen3-VL-A3B-30B (Yang et al., 2025)				
Qwen3-VL-A3B-30B	59.32	64.08	54.29	61.33
MH-MoE	64.85	65.92	58.82	50.50
Metis-HOME	65.81	68.45	61.40	65.30
I^2MoE	64.27	68.71	53.46	**68.00**
CL-MOE	65.31	73.30	48.69	48.65
MoME	66.78	70.64	56.39	56.75
CoGR-MoE (Ours)	**68.96**	**74.54**	**64.61**	62.83

Table 2: Comparison of MoE-based LLaVA-1.5-7B (Liu et al., 2024) on five standard VL benchmarks in VMCBench, reporting accuracy across VQAv2(Goyal et al., 2017), GQA(Hudson and Manning, 2019), VizWiz(Gurari et al., 2018), ScienceQA(Lu et al., 2022a), MMVet(Yu et al., 2024), and MMStar(Chen et al., 2024).

Method	VQAv2	GQA	VizWiz	ScienceQA	MMVet	MMStar
LLaVA-1.5-7B	66.7	72.6	68.6	64.3	54.0	34.2
MH-MoE	83.6	75.3	77.0	77.4	64.0	39.9
I^2MoE	83.3	81.1	82.4	74.6	**66.7**	47.8
CL-MOE	79.2	78.2	80.9	**83.0**	60.4	42.8
MoME	75.7	81.2	81.0	80.4	63.3	48.1
Metis-HOME	82.7	**83.2**	78.5	80.5	64.7	50.4
CoGR-MoE (Ours)	**88.5**	**83.2**	**84.8**	77.4	64.3	**52.0**

4.2 Analysis

As shown in Table 1, CoGR-MoE yields the highest overall accuracy on two backbones, reaching 63.25 on MOE-LLaVA and 68.96 on Qwen3-VL-A3B-30B. It shows notable gains in Perspective and Transformative tasks, highlighting its strength in handling partial-visibility and spatial transformations. However, improvements on the Others subset remain limited compared with I^2MoE and MoME. CoGR-MoE demonstrates superior performance across the VMCBench in Table 2, particularly excelling in tasks like VQAv2 and VizWiz. However, its performance on the ScienceQA and MMVet tasks shows less improvement, lagging behind CL-MoE and Metis-HOME.

The ablation results in Table 3 highlight the contribution of each component to the overall performance of CoGR-MoE. On almost all datasets, the Full model still performs better. Notably, the w/o unc_j variant outperforms the Full model on ScienceQA which involves more direct reasoning, achieving a score of 78.0, compared to the Full model's 77.6. In addition, the removal of s_a and L_{distill}, as well as the Prompt-only variant, had the largest negative impact on accuracy, causing a significant drop in performance across multiple datasets.

In Table 4, the best results occur at $n = 8$ when n

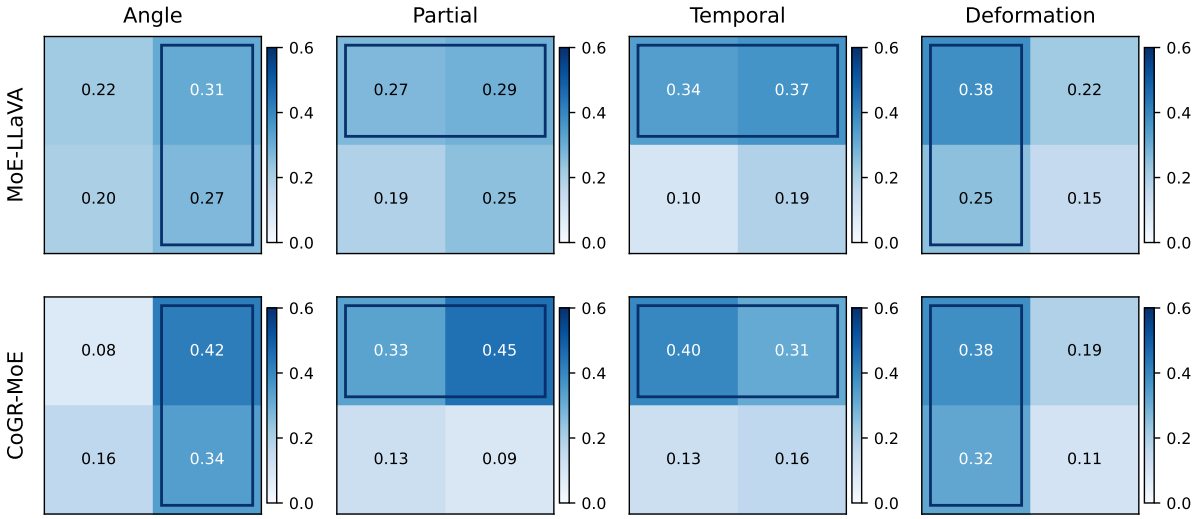

Figure 3: Each heatmap summarizes expert routing behavior of MoE-LLaVA and CoGR-MoE under four subtasks in MRAG-Bench. Each cell corresponds to a single expert. Color intensity indicates the probability of an expert being selected into the Top-K routing set. Darker colors correspond to experts that are more frequently selected under the given task type. Compared to MoE-LLaVA, CoGR-MoE exhibits more concentrated and task-consistent routing patterns.

changes, reaching an accuracy of 71.6 and a Sim of 0.48, noticeably higher than smaller settings. When varying K under $n = 8$, activating $K = 2$ yields the highest accuracy of 74.5 and the strongest Sim of 0.51, whereas larger K values lead to a decline.

4.3 Discussion

CoGR-MoE excels because it learns to select experts based on correct semantic direction and consistently assigns similar questions to the same experts. Figure 3 shows that consistent routing of similar questions promotes reuse of experts aligned with content requirements. To further support fine-grained distinctions among answer options, option-specific cues are introduced to dynamically reweight the shared Top-K experts. This enables option-level discrimination without disrupting established expert roles, as illustrated in Figure 4. Even when the initially selected experts provide weakly discriminative scores, reweighting can still amplify relative differences across options. This effect is most pronounced in Perspective and Transformative tasks involving viewpoint changes or partial visual evidence. In contrast, gains on VQAv2 are modest, while ScienceQA shows more unstable effects due to higher uncertainty.

However, CoGR-MoE underperforms in the task like Other because it often fails to capture the subtle differences in cross-modal interactions. The core limitation lies in that its routing strategy is primarily guided by the overall semantic direction of the input, rather than explicitly modeling fine-grained cross-modal interactions. Concept-guided gating applies a consistent expert-selection strategy aligned with the overall semantic direction of the input, whereas interaction-driven gating adapts expert selection to the instance-specific cross-modal interactions required by each sample. In tasks that demand highly dynamic and complex multimodal interactions, routing based primarily on semantic alignment may be insufficient to capture the full range of instance-specific dependencies. In contrast, I²MoE excels by using interaction-specific experts for each modality pair, allowing more precise expert selection in tasks with complex intermodal relationships. Similarly, MoME performs well by dynamically activating expert pools specialized for visual and textual tasks.

Prompt-only and the w/o s_a variants underperform because inference-time prompting or unguided routing does not update the router. By contrast, cue-based training internalizes attribute–expert associations through repeated gradient reinforcement. In addition, removing the contrastive loss disrupts CoGR-MoE's ability to generate more discriminative expert representations. This is especially important in tasks like VQAv2 and GQA, which require fine-grained reasoning. In the case of ScienceQA, many challenging examples arise from reasoning complexity or from

Table 3: Ablation results for the components built upon the MoE-LLaVA backbone.

Method	VQAv2	GQA	VizWiz	ScienceQA	MMVet	MMStar	MRAG
w/o s_a	80.4	76.8	80.4	70.1	63.2	46.3	57.9
w/o s_j	84.6	82.3	82.7	76.1	64.1	50.9	60.8
w/o unc_j	85.2	80.8	83.5	**78.0**	62.5	50.8	61.0
w/o $L_{contrast}$	87.8	84.6	82.1	73.0	60.4	49.5	59.3
w/o $L_{distill}$	83.3	82.4	79.8	75.8	63.2	50.4	58.9
Prompt-only	81.5	77.6	79.4	65.7	63.6	45.1	57.3
Full	**88.2**	**85.1**	**85.8**	77.6	**65.7**	**53.4**	**63.3**

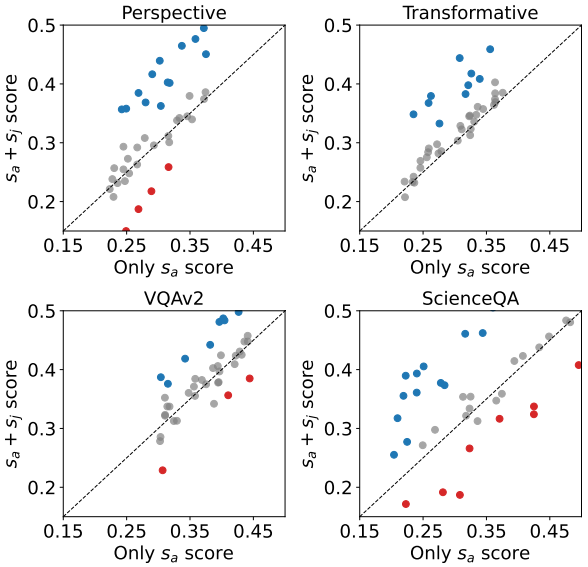

Figure 4: Comparison of answer scoring with and without option-level reweighting. The top row shows subtasks Perspective and Transformative in MRAG-Bench and the bottom row shows VQAv2 and ScienceQA in VMCBench, each with 40 randomly sampled instances. The x-axis shows scores computed using only s_a, whereas the y-axis shows scores of $s_a + s_j$. Blue and red points indicate notable score increases and decreases, respectively, while gray points indicate minor changes.

images that lack clear visual evidence for the correct answer. Uncertainty-based weighting reduces the gradient updates for cue-weak but informative examples, while allowing easier, cue-dominated examples to dominate the gradient updates during optimization. Under this training scheme, the Full model places less emphasis on cue-weak examples that may be more instructive, which hinder its ability to acquire fine-grained reasoning skills.

Performance peaks at $n = 8$, indicating that a balanced expert pool provides sufficient diversity for specialization. Smaller configurations lack capacity for experts to develop distinct semantic roles, while overly large pools dilute expert utilization. Under a fixed $n = 8$, the optimal setting occurs at $K = 2$, where experts can contribute complementary semantic information while maintaining focused routing. Larger K values introduce semantic dilution, as aggregating many weakly relevant experts reduces both accuracy and semantic alignment.

Table 4: Effect of the number of experts n and activated experts K on accuracy and Sim on VMCBench. When varying n, the number of activated experts is fixed to $K = 1$; when varying K, the total number of experts is fixed to $n = 8$.

Method	n	Acc	Sim	K	Acc	Sim
CoGR-MoE	1	68.2	0.36	1	71.6	0.48
	2	64.5	0.27	2	**74.5**	**0.51**
	4	59.1	0.43	3	68.6	0.38
	8	**71.6**	**0.48**	4	64.2	0.44
	16	55.4	0.30	5	62.0	0.31

5 Conclusion

In this paper, we propose CoGR-MoE, a concept-guided MoE framework that mitigates routing inconsistency by injecting the correct answer's semantic direction into router. Furthermore, the option-aware weighting mechanism dynamically reweights the Top-K experts, enhancing the model's ability to discriminate. Experiments on multiple multimodal benchmarks demonstrate the superiority of CoGR-MoE in improving accuracy across diverse VQA tasks. Ablation studies further confirm the effectiveness of each component. Future work will focus on extending CoGR-MoE's routing mechanism to better account for varying cross-modal patterns and task complexities.

6 Limitations

Despite its effectiveness, CoGR-MoE has several limitations. First, it relies on LLM-generated semantic cues during training, introducing additional computational cost and dependency on external language models, which may limit scalability for large-scale or frequent retraining settings.

Second, although the agreement-variance mechanism alleviates unstable or noisy cues, the model may still be sensitive to systematic biases or hallucinations from the LLM, particularly for visually ambiguous or fine-grained concepts.

Finally, our evaluation is limited to multiple-choice visual question answering benchmarks, and the effectiveness of the proposed mechanisms for open-ended generation or other multimodal reasoning tasks remains to be explored.

References

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. 2025. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 182–210, Vienna, Austria. Association for Computational Linguistics.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, page 1–20.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Anzhe Cheng, Shukai Duan, Shixuan Li, Chenzhong Yin, Mingxi Cheng, Heng Ping, Tamoghna Chattopadhyay, Sophia I Thomopoulos, Shahin Nazarian, Paul Thompson, and Paul Bogdan. 2025a. Ermoe: Eigen-reparameterized mixture-of-experts for stable routing and interpretable specialization. *Preprint*, arXiv:2511.10971.

YuJu Cheng, Yu-Chu Yu, Kai-Po Chang, and Yu-Chiang Frank Wang. 2025b. Serial lifelong editing via mixture of knowledge experts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30888–30903, Vienna, Austria. Association for Computational Linguistics.

Haojie Duanmu, Xiuhong Li, Zhihang Yuan, Size Zheng, Jiangfei Duan, Xingcheng Zhang, and Dahua Lin. 2025. Mxmoe: Mixed-precision quantization for moe with accuracy and performance co-design. *Preprint*, arXiv:2505.05799.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *Preprint*, arXiv:1802.08218.

Xing Han, Hsing-Huan Chung, Joydeep Ghosh, Paul Pu Liang, and Suchi Saria. 2025. Guiding mixture-of-experts with temporal multimodal interactions. *Preprint*, arXiv:2509.25678.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. 2025. Clmoe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. *arXiv preprint arXiv:2503.00413*.

Shaohan Huang, Xun Wu, Shuming Ma, and Furu Wei. 2024. Mh-moe: Multi-head mixture-of-experts. *Preprint*, arXiv:2411.16205.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Comput. Surv.*, 57(10).

Xiaohan Lan, Fanfan Liu, Haibo Qiu, Siqi Yang, Delian Ruan, Peng Shi, and Lin Ma. 2025. Metis-home: Hybrid optimized mixture-of-experts for multimodal reasoning. *arXiv preprint arXiv:2510.20519*.

Minh Le, An Nguyen, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Van Ngo, and Nhat Ho. 2025. Mixture of experts meets prompt-based continual learning. *Preprint*, arXiv:2405.14124.

Jing Li, Zhijie Sun, Dachao Lin, Xuan He, Binfan Zheng, Yi Lin, Rongqian Zhao, and Xin Chen. 2025a. Expert-token resonance moe: Bidirectional routing with efficiency affinity-driven active selection. *Preprint*, arXiv:2406.00023.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, Baotian Hu, and Min Zhang. 2025b. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *Preprint*, arXiv:2511.12609.

Zhongyang Li, Ziyue Li, and Tianyi Zhou. 2025c. R2-t2: Re-routing in test-time for multimodal mixture-of-experts. *Preprint*, arXiv:2502.20395.

Zhongyang Li, Ziyue Li, and Tianyi Zhou. 2025d. Routing manifold alignment improves generalization of mixture-of-experts llms. *Preprint*, arXiv:2511.07419.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2022b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *Preprint*, arXiv:2110.13214.

Siyuan Mu and Sen Lin. 2025. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *Preprint*, arXiv:2503.07137.

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. 2025. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. In *The Thirteenth International Conference on Learning Representations*.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. 2025. Probing semantic routing in large mixture-of-expert models. *Preprint*, arXiv:2502.10928.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. 2024a. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Preprint*, arXiv:2407.12709.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. 2024b. Mome: Mixture of multimodal experts for generalist multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vinay Kumar Verma, Shreyas Sunil Kulkarni, Happy Mittal, and Deepak Gupta. 2025. Moemoe: Question guided dense and scalable sparse mixture-of-expert for multi-source multi-modal answering. *Preprint*, arXiv:2503.06296.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Yujin Han, Zhekai Chen, Jiayu Wang, Difan Zou, Xihui Liu, Yingya Zhang, Yu Liu, and Hongming Shan. 2025. Routing matters in moe: Scaling diffusion transformers with explicit routing guidance. *Preprint*, arXiv:2510.24711.

Qiong Wu, Zhaoxi Ke, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. 2025. Routing experts: Learning to route dynamic experts in multi-modal large language models. *Preprint*, arXiv:2407.14093.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. 2025. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *arXiv preprint arXiv:2505.19190*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. 2025. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. *Preprint*, arXiv:1511.03416.

A Training Setup

We fine-tune Qwen3-VL-30B, MoE-LLaVA, and LLaVA-1.5-7B on the IconQA, A-OKVQA, and Visual7W training sets, where 5,000 samples are randomly selected from each dataset. LLaVA-1.5-7B is converted to an MoE model by replacing dense MLP blocks with 8-expert Top-2 gated MoE layers. For Qwen3-VL-30B and MoE-LLaVA, we follow their official MoE configurations as specified in the original model releases.

We use the AdamW optimizer with a base learning rate of 1×10^{-4} and a per-GPU batch size of 4. Training is conducted on 4 NVIDIA A100 GPUs with gradient accumulation over 4 steps, resulting in an effective batch size of 64. A cosine learning rate scheduler (CosineAnnealingLR) is applied with a minimum learning rate of 1×10^{-6}

and a warmup period of 100 steps. Gradient clipping with a maximum norm of 1.0 is used to stabilize training. In addition, sample-level gradient reweighting is applied at the routing layer based on $Agr(j)$ and $Var(j)$, with weights clipped to the range $[0.1, 1.0]$.

Compared to a standard Top-K MoE architecture, our approach retains the same gating scheme and does not increase the number of expert forward passes during training or inference, introducing only a lightweight semantic term in the routing logits. At inference time, the model exhibits the same inference-time complexity and memory footprint as the corresponding baseline MoE models.

Semantic cues are generated offline using a large language model, with one-time generation per question prior to training. The generated cues are cached and reused throughout training and evaluation. As a result, no LLM calls are made during training iterations or inference, and the LLM is not part of the training or inference loop. All reported results are averaged over three independent runs.

B Additional Experiments

B.1 Case Study

Figure 5 presents an example from the Temporal category of MRAG-Bench, comparing the behaviors of CoGR-MoE and I^{2}MoE. The image depicts a large neoclassical building under construction, while the question asks for the name of the architecture once construction is finished. Among the four candidate answers, the correct choice is Pantheon.

We compare the routing behavior and decision outcomes of MH-MoE and CoGR-MoE on this example. I^{2}MoE exhibits a relatively diffuse routing pattern, with expert activations distributed more evenly across the selected experts. As a result, the aggregated representations formed for different answer options are weakly differentiated, leading the model to assign a higher similarity score to an incorrect option. In contrast, CoGR-MoE shows a more structured routing behavior. Although the same Top-K experts are selected, CoGR-MoE applies option-specific reweighting over these experts, resulting in distinct aggregated representations for each answer choice.

This difference in expert CoGR is reflected in the final option scores. CoGR-MoE substantially increases the score of the correct answer while suppressing the scores of distractors, thereby enlarging the margin between the correct option and all alternatives. I^{2}MoE , by contrast, fails to separate the correct answer from visually or semantically similar distractors.

B.2 Uncertainty Study

To evaluate the contribution of stability-aware sample weighting, we disable the variance component in uncertainty estimation. As a result, gradient weighting depends solely on agreement, without suppressing unstable or noisy probes. All options therefore receive similar training strength independent of textual consistency. To further isolate the role of agreement-based semantic reliability, we construct an Only-Variance baseline by removing the agreement term. In this setting, uncertainty is determined solely by probe variation, without incorporating agreement information. This variant treats all semantic cues as equally plausible regardless of their semantic alignment with the image–question pair. It keeps the ability to detect probe inconsistency (larger variance $\rightarrow$ higher uncertainty) but removes any information about whether the probes actually align with the image–question pair.

Across all seven benchmarks, removing uncertainty consistently lowers accuracy as shown in Table 7. The largest drops appear on VQAv2, with a decrease of 4.1 percent. More moderate declines occur on ScienceQA with 1.5 percent and MMVet with 0.8 percent. The accuracy of Only-Variance falls between that of w/o uncertainty and Full CoGR. However, on ScienceQA, Only-Variance obtains 75.4 percent, slightly lower than the 75.9 percent of w/o uncertainty, and on MMVet, Only-Variance reaches 62.8 percent, also the lowest.

The largest performance drops on VQAv2 and MRAG stem from the fact that they contain highly diverse, semantically overlapping answer options and substantial linguistic ambiguity, which makes them particularly sensitive to noisy or unstable probe signals. Without uncertainty, noisy probes receive equal training strength, contaminating expert routing, distorting the semantic direction for contrastive alignment, and amplifying option-level confusion. ScienceQA, however, is largely unaffected by this removal because its questions are highly structured and semantically unambiguous. As a result, the model relies far less on statement stability for distinguishing candidate answers, leading to only minimal degradation when uncertainty is removed.

On ScienceQA, the Only-Variance variant under-

Table 5: Performance comparison on the **Perspective** categories of MRAG-Bench.

Methods	Angle	Partial	Scope	Occlusion
MOE-LLaVA				
MOE-LLaVA	62.42	54.88	60.19	47.65
MH-MoE	68.94	**69.92**	**73.53**	66.74
Metis-HOME	68.77	62.68	65.24	58.47
I^2MoE	65.49	61.76	64.87	59.53
CL-MOE	65.69	63.41	58.82	66.67
MoME	68.95	69.42	73.24	65.57
CoGR-MoE (Ours)	**70.29**	**69.92**	67.35	**68.81**
Qwen3-VL-A3B-30B				
Qwen3-VL-A3B-30B	62.80	61.94	63.78	67.81
MH-MoE	71.01	**73.42**	57.90	61.33
Metis-HOME	64.78	68.54	69.83	70.65
I^2MoE	69.53	67.29	72.61	65.42
CL-MOE	72.61	72.92	73.37	74.30
MoME	74.18	72.54	66.86	68.96
CoGR-MoE (Ours)	**77.78**	71.26	**73.61**	**75.63**

performs because questions are highly structured, so variance mainly reflects random fluctuations in statement scoring rather than meaningful semantic stability. On MMVet's instruction- and reasoning-style questions, the performance drop is even more pronounced. Correct answers are often longer, multi-aspect explanations, which naturally induce higher variance across statements, whereas plausible but incorrect distractors can appear more uniform and thus exhibit lower variance. A variance-only scheme therefore tends to down-weight precisely those samples that require nuanced reasoning, while relatively up-weighting simpler but incorrect options. In contrast, the Full model formulation combines variance with agreement-based alignment, allowing the model to discount probes that are stable yet semantically misaligned.

B.3 Prompt Ablation Study

To isolate prompt effects from the CoGR-MoE architecture itself, we compare two cues-generation settings under an otherwise identical training and evaluation pipeline. We consider the following two prompts for the LLM-based cues generator:

- **Full Prompt.** This is the task-aware prompt used in our main experiments. It explicitly instructs the LLM to (i) distinguish between positive-cues and negative-cues, and (ii) attend to task characteristics (e.g., object recognition, counting, OCR-like text reading, and relational reasoning).

- **Minimal Prompt.** This ablated version removes all task-type hints and high-level reasoning instructions. The LLM is only asked to generate a small set of positive-cues and negative-cues for each answer option. No additional guidance about problem type, reasoning strategy, or probe diversity is provided.

Across the seven benchmarks, using the Minimal Prompt instead of the Full Prompt leads to only small drops on the standard VQA datasets. The decrease is very small on VQAv2, where accuracy falls from 88.5 to 87.8, and similarly small on GQA, where it goes from 83.2 to 82.6. Larger differences appear on more complex reasoning datasets: MMVet from 64.3 to 60.9, and MMStar from 52.0 to 47.5. The decline is most pronounced on reasoning-oriented and instruction-following tasks.

The prompt ablation highlights a clear task-dependent pattern. On perception-oriented VQA datasets such as VQAv2 and GQA, replacing the full task-aware prompt with a minimal cue prompt leads to only minor degradation, suggesting that these benchmarks rely mainly on local visual evidence that simple support/reject cues can already capture.

Table 6: Performance comparison on the **Transformative** and **Others** categories of MRAG-Bench.

Methods	Temporal	Deformation	Incomplete	Biological	Others
MOE-LLaVA					
MOE-LLaVA	47.65	53.92	47.62	53.92	51.67
MH-MoE	**70.47**	42.16	41.22	52.94	40.83
Metis-HOME	58.59	50.28	40.63	54.90	55.75
I^2MoE	54.94	57.69	32.06	58.80	60.83
CL-MOE	63.12	47.06	39.22	56.86	59.94
MoME	69.30	54.71	39.82	**59.80**	**64.06**
CoGR-MoE (Ours)	68.46	**58.80**	**49.18**	57.75	58.67
Qwen3-VL-A3B-30B					
Qwen3-VL-A3B-30B	57.36	51.85	**52.02**	55.94	61.33
MH-MoE	71.46	58.88	34.37	65.78	50.50
Metis-HOME	74.26	62.67	51.94	64.32	65.30
I^2MoE	62.73	60.41	42.22	60.84	**68.00**
CL-MOE	71.46	47.12	41.27	56.92	48.65
MoME	73.05	56.92	41.37	63.26	56.75
CoGR-MoE (Ours)	**74.84**	**65.21**	51.62	**66.75**	62.83

Table 7: Ablation accuracy on uncertainty components.

Method	**VQAv2**	**GQA**	**VizWiz**	**ScienceQA**	**MMVet**	**MMStar**	**MRAG**
w/o unc	82.4	81.1	83.0	75.9	63.5	49.7	59.5
Only-Variance	85.8	82.5	84.2	75.4	62.8	50.9	62.7
Full	**88.5**	**83.2**	**84.8**	**77.4**	64.3	**52.0**	**63.3**

In contrast, reasoning-oriented and instruction-following tasks, including ScienceQA, MMVet, MMStar, and MRAG show larger drops. These datasets require understanding multi-step instructions and applying task-specific constraints, and the full prompt provides essential guidance for generating cues that reflect such reasoning steps. Without this guidance, the cues become more superficial, reducing the quality of semantic anchoring and routing. Overall, CoGR-MoE remains robust on perception-heavy tasks but benefits more from structured prompting when deeper reasoning is required.

B.4 Routing Sharpness and Variance Study

To evaluate whether CoGR-MoE achieves more semantically aligned and decisive routing than MoE-LLaVA, routing is explicitly enhanced toward experts aligned with the correct semantic direction. Qwen-VL is not included here because its MoE layer uses 128 experts with Top-8 routing, under which the relative differences in gating distributions become extremely small and difficult to interpret. Expert-gating distributions are collected on MRAG-Bench, where each test sample is annotated with a semantic category (e.g., Angle, Occlusion, Deformation, Biological). Two complementary metrics are used:

Routing sharpness is defined as the difference between the average gating weight of the selected experts and that of the unselected experts. Formally, letting $T(x)$ denote the Top-K experts selected for sample x:

$$\begin{aligned}\text{Sharp}(x) = &\frac{1}{|T(x)|}\sum_{i\in T(x)} g_i(x) \\ &- \frac{1}{E-|T(x)|}\sum_{i\notin T(x)} g_i(x),\end{aligned} \tag{12}$$

where E is the total number of experts. A higher value indicates stronger preference for the selected experts.

Routing variance is defined for each semantic category c as:

$$\text{Var}_c = \frac{1}{E}\sum_{i=1}^{E}\text{Var}_{x\in c}\big(g_i(x)\big), \tag{13}$$

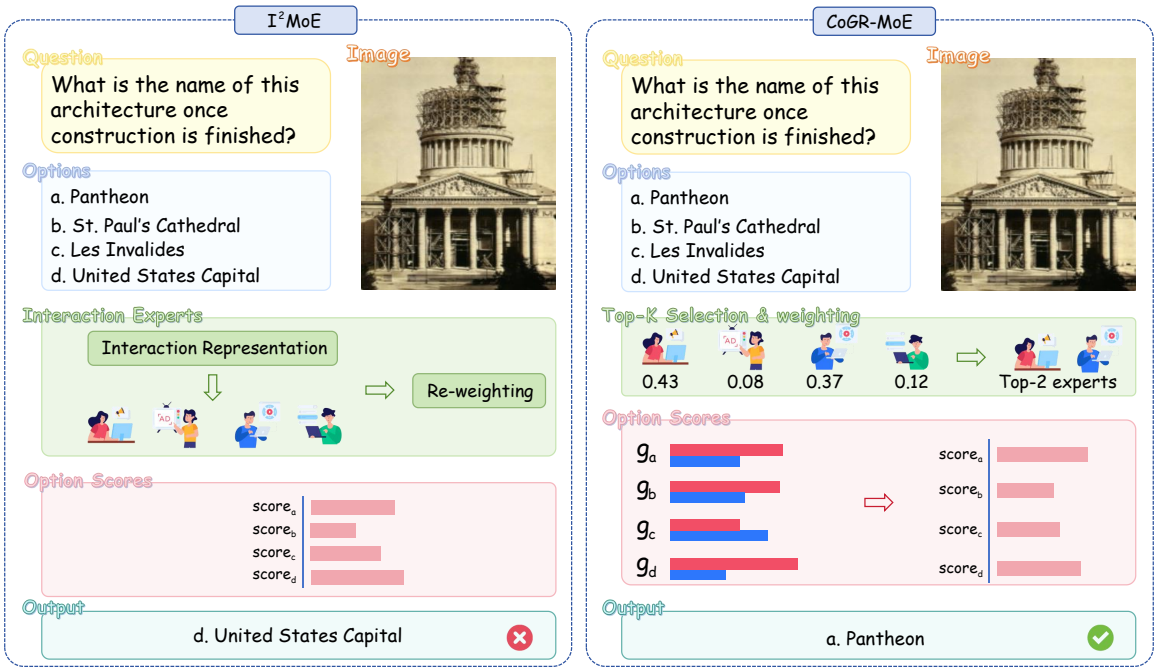

Figure 5: Qualitative comparison between CoGR-MoE and I^{2}MoE on a temporal question from MRAG-Bench. Although both models process the same image–question pair, CoGR-MoE produces more discriminative option scores by leveraging structured expert uCoGR, leading to the correct prediction.

Table 8: Performance comparison of CoGR-MoE under different prompt configurations. The Full Prompt provides detailed visual cue guidance, while the Minimal Prompt only generates simple postive and negative cues. All other components of CoGR-MoE remain identical.

Method	VQAv2	GQA	VizWiz	ScienceQA	MMVet	MMStar	MRAG
Minimal Prompt	87.8	82.6	82.1	74.3	60.9	47.5	59.6
Full Prompt	**88.5**	**83.2**	**84.8**	**77.4**	**64.3**	**52.0**	**63.3**

which quantifies the variability of expert-gating distributions among samples within the same semantic group, with lower variance indicating more stable and semantically consistent routing. To improve readability, we report routing variance scaled by a factor of 10 without affecting relative comparisons.

As shown in Table 9 and Table 10, CoGR-MoE produces consistently sharper and more stable routing than MoE-LLaVA. For routing sharpness, CoGR-MoE on MoE-LLaVA increases the overall score from 0.12 to 0.23. Stronger improvements appear in categories such as Angle and Partial. For routing variance, CoGR-MoE substantially reduces intra-category variability. On MoE-LLaVA, the overall variance drops from 0.46 to 0.31, with larger reductions in Partial and Temporal.

The improvements in routing sharpness and variance stem from CoGR-MoE's ability to explicitly steer the router toward experts aligned with the correct semantic direction. Rather than relying on unconstrained gating logits, CoGR-MoE builds a semantic anchor from positive and negative cues and injects it as a learnable signal, nudging samples of the same semantic category toward a consistent subset of experts. The probe-aligned contrastive objective further strengthens this direction, while uncertainty-aware weighting suppresses noisy gradients that could destabilize routing.

C Theoretical Analysis

C.1 Bounded Routing Perturbation

Proposition. Let $z_{\text{base}} \in R^E$ denote the original gating logits over E experts, and let the semantic vector b_{CoGR} satisfy $\|b_{\text{CoGR}}\|_\infty \leq C$ for some constant $C > 0$. For any bounded scalar λ, the semantic-guided routing

$$g^T = \text{softmax}\big(z_{\text{base}} + \lambda b_{\text{CoGR}}\big) \qquad (14)$$

Table 9: Comparison of routing sharpness and routing variance on the **Perspective** categories.

Methods	**Overall**	Angle	Partial	Scope	Occlusion
Routing Sharpness					
MOE-LLaVA	0.12	0.08	0.06	0.12	0.08
CoGR-MoE (MOE-LLaVA)	0.23	0.26	0.28	0.26	0.20
Routing Variance					
MOE-LLaVA	0.46	0.35	0.53	0.46	0.31
CoGR-MoE (MOE-LLaVA)	0.32	0.19	0.22	0.30	0.24

Table 10: Comparison of routing sharpness and routing variance on the **Transformative** and **Other** categories.

Methods	Temporal	Deformation	Incomplete	Biological	Others
Routing Sharpness					
MOE-LLaVA	0.21	0.13	0.07	0.13	0.18
CoGR-MoE (MOE-LLaVA)	0.20	0.20	0.24	0.18	0.20
Routing Variance					
MOE-LLaVA	0.65	0.43	0.24	0.46	0.72
CoGR-MoE (MOE-LLaVA)	0.47	0.33	0.21	0.35	0.58

constitutes a bounded and Lipschitz-continuous perturbation of the original routing distribution $\mathrm{softmax}(z_{\text{base}})$. In particular, it does not introduce additional extrema nor destabilize the relative ordering of experts.

Proof. The softmax function is Lipschitz-continuous with respect to its input logits under the ℓ_∞ norm. That is, there exists a constant $L > 0$ such that for any $z, z' \in R^E$,

$$\|\mathrm{softmax}(z) - \mathrm{softmax}(z')\|_1 \leq L\|z - z'\|_\infty. \tag{15}$$

In the semantic-guided routing, the perturbation applied to the original logits is

$$(z_{\text{base}} + \lambda b_{\text{CoGR}}) - z_{\text{base}} = \lambda b_{\text{CoGR}}. \tag{16}$$

Since $\|b_{\text{CoGR}}\|_\infty \leq C$, we have

$$\|(z_{\text{base}} + \lambda b_{\text{CoGR}}) - z_{\text{base}}\|_\infty \leq \lambda C. \tag{17}$$

Combining the above inequalities, the induced change in the routing distribution is bounded by $L\lambda C$. Therefore, the semantic enhancement introduces a controlled directional shift in logit space rather than an unbounded deformation, preserving routing stability and preventing expert collapse.

C.2 Shared Top-K Consistency

Let z_{base} denote the base gating logits produced by the router for an image–question pair (I, Q), and let TopK denote the set of experts selected according to z_{base}. For each answer option j, option-specific logits over the shared Top-K experts are defined as

$$\mathrm{logits}_{\text{top}}(j) = z_{\text{base}}[\mathrm{TopK}] + \lambda s_j[\mathrm{TopK}], \tag{18}$$

where s_j denotes the semantic signal derived from the option text. The corresponding gating distribution is given by

$$g_j = \mathrm{softmax}\big(\mathrm{logits}_{\text{top}}(j)\big), \tag{19}$$

and the aggregated representation for option j is computed as

$$\tilde{h}_j = \sum_{i \in \mathrm{TopK}} g_j(i)\, h_i. \tag{20}$$

Since the Top-K expert set is determined solely by z_{base}, it is identical for all answer options and independent of the option index j. Consequently, routing decisions are shared across options and depend only on the input (I, Q). s_j influences the model exclusively through reweighting within this shared expert set, without altering the routing outcome.

As all option representations $\tilde{h}_j$ are linear combinations of the same expert outputs, they lie in a common expert subspace. This design ensures comparability across options and prevents option-conditional routing drift. In contrast, allowing each option to independently select its own Top-K experts would entangle routing with option identity, leading to duplicated expert computation and destabilized expert specialization.

C.3 Bounded Gradient Reweighting

Proposition. Let the main training objective be defined as

$$L_{main} = \sum_j \frac{1}{1 + unc_j} \mathrm{CE}(\mathrm{score}(j), y_j). \quad (21)$$

where $\mathrm{CE}_j(\theta)$ denotes the cross-entropy loss associated with option j, and $unc_j \geq 0$ is the corresponding uncertainty-based weight. This formulation induces a bounded rescaling of per-option gradients and does not introduce uncontrolled gradient amplification.

Proof. Taking the gradient of L_{main} with respect to model parameters θ yields

$$\nabla_\theta L_{main} = \sum_j unc_j \, \nabla_\theta \mathrm{CE}_j(\theta). \quad (22)$$

Applying the triangle inequality, we obtain

$$\begin{aligned} \|\nabla_\theta L_{main}\| &\leq \sum_j unc_j \, \|\nabla_\theta \mathrm{CE}_j(\theta)\| \\ &\leq \big(\max_j unc_j\big) \sum_j \|\nabla_\theta \mathrm{CE}_j(\theta)\|. \end{aligned} \quad (23)$$

Therefore, uncertainty-aware weighting amounts to a bounded linear rescaling of per-option gradients. As long as unc_j is bounded, the overall gradient magnitude remains controlled, ensuring that uncertainty-based reweighting does not destabilize optimization but only modulates the relative influence of different options during training.

C.4 Contrastive-Alignment Consistency

Proposition. Let the semantic alignment metric be defined as

$$\mathrm{Sim} = \cos(h_{\mathrm{TopK}}, s_a), \quad (24)$$

where h_{TopK} denotes the aggregated representation of the routed Top-K experts and s_a is the semantic direction corresponding to the correct answer. Consider the cue-guided contrastive objective used in Eq. (9), which encourages the aggregated representation of the correct option to align with positive semantic cues while pushing incorrect option representations away from negative cues. Then, minimizing the contrastive loss induces gradient updates on the shared expert representations that are directionally consistent with increasing the semantic alignment metric Sim.

Proof. The cosine similarity measures the angular alignment between two vectors in representation space. Without loss of generality, we assume $\|s_a\| = 1$, since normalization does not affect directional analysis. For a representation h, the cosine similarity with respect to s_a can be written as

$$\cos(h, s_a) = \frac{h^\top s_a}{\|h\|}. \quad (25)$$

Taking the gradient with respect to h yields

$$\nabla_h \cos(h, s_a) = \frac{s_a}{\|h\|} - \cos(h, s_a) \frac{h}{\|h\|^2}. \quad (26)$$

The first term promotes rotation of h toward the semantic direction s_a, while the second term removes the component aligned with h itself, preventing trivial norm inflation. As a result, minimizing a negative cosine similarity term $-\cos(h, s_a)$ induces gradient updates that rotate the representation toward s_a while keeping its norm controlled.

In our setting, the contrastive loss in Eq. (9) is defined over option-specific aggregated representations, including the correct-option representation $\tilde{h}_{\mathrm{correct}}$ and a pooled incorrect-option representation $\tilde{h}_{\mathrm{wrong}}$. Both representations are formed by reweighting the same shared Top-K expert outputs. Consequently, gradients from the contrastive objective propagate to the shared expert representations that constitute h_{TopK}.

The positive term in the contrastive loss aligns $\tilde{h}_{\mathrm{correct}}$ with the positive semantic cues, which are constructed to be consistent with the semantic direction s_a. This induces gradient updates on the selected experts that rotate their representations toward s_a. Meanwhile, the negative term discourages alignment of incorrect-option representations with incompatible semantic cues, further suppressing expert activations that conflict with the correct semantic direction.

Therefore, although the contrastive objective does not explicitly maximize Sim, its gradient induces a consistent rotational pressure on the shared

expert representations toward the correct semantic direction. As training proceeds, this alignment-consistent gradient flow increases the expected cosine similarity between h_{TopK} and s_a, corresponding to an increase in the semantic alignment metric Sim.

D Prompt Listing

D.1 Full Prompt

```
You are a visual evidence aligner.

Given an image I, a question Q, and a
↪ set of candidate answer options,
your goal is to generate decision-level
↪ semantic cues for each option.

For each option a, produce two sets of
↪ cues:
- positive-cues: observable visual
↪ evidence that should be present if a
↪ is correct.
- negative-cues: observable evidence
↪ that would contradict a if present.

Before generating cues, identify the
↪ main task type(s) involved in the
↪ question
(multiple types may apply), and tailor
↪ the cues accordingly.

Task categories include:
- Perception
- Counting
- Spatial
- OCR / Text
- Commonsense & Knowledge
- Reasoning

[Category-Specific Guidance]

Perception:
Focus on directly observable visual
↪ attributes, such as object parts,
↪ colors, materials,
textures, or the presence or absence of
↪ specific entities. Cues should be
↪ localized
and visually verifiable.

Counting:
Focus on explicit countable instances or
↪ visual anchors that support a
↪ specific quantity.
Negative cues should highlight visible
↪ evidence that contradicts an
↪ incorrect count.

Spatial:
Focus on relative spatial relationships
↪ between entities, such as left/right,
↪ above/below,
inside/outside, distance, orientation,
↪ or occlusion relationships.

OCR / Text:
Focus on visible text or symbols in the
↪ image. Positive cues should describe
↪ readable text
or character patterns together with
↪ coarse location information. Avoid
↪ hallucinating text;
if the text is uncertain, indicate low
↪ confidence rather than fabricating
↪ content.

Commonsense & Knowledge:
Use only cues that are directly
↪ supported by observable visual
↪ evidence.
Do not rely on external knowledge as
↪ decisive evidence when generating
↪ cues.

Reasoning:
Describe minimal sets of observable
↪ premises together with a short rule
↪ or relation
that supports the option. Avoid long
↪ chains of reasoning and do not
↪ introduce
external knowledge beyond what is
↪ visible or stated in the question.

[General Principles]
- Cues should be concrete, specific, and
↪ verifiable from the image.
- Prefer minimal but discriminative
↪ evidence.
- Avoid subjective or non-observable
↪ attributes.
- If evidence is weak or uncertain,
↪ reflect this by lowering confidence
↪ rather than inventing details.

Return the positive-cues and
↪ negative-cues for each option in a
↪ structured format.
```

D.2 Minimal Prompt

```
You are given an image, a question, and
↪ multiple candidate answer options.

For each option a, generate:
- positive-cues: a small set of visible
↪ cues that must be present in the
↪ image if the option were correct.
- negative-cues: visible cues that, if
↪ present, would contradict the
↪ option.

Cues should be concrete, observable, and
↪ directly verifiable from the image.
Do not rely on external knowledge or
↪ subjective descriptions.
If evidence is weak or uncertain, lower
↪ the confidence rather than inventing
↪ details.

Output a concise JSON object containing,
↪ for each option, its positive-cues
↪ and negative-cues.
```