

Appendix

This document supplements the main paper as follows:

- Sec. A provides detailed proofs for the proposed *frontier search*;
- Sec. B details additional experimental setups;
- Sec. C provides additional comparisons with baselines across different *expert skipping* ratios and MLLMs;
- Sec. D presents visual question answering examples across methods;
- Sec. E reports ablations on the number of grid points in *frontier search*;
- Sec. F shows ablations on the number of samples used for calibration and search;
- Sec. G analyzes expert redundancy *w.r.t.* tokens across modalities.

A. Proofs

In this section, we first provide complete proofs of the correctness and time complexity for our frontier search (Prop. 1). We then prove that the optimal thresholds lie on the frontier (Prop. 2).

Lemma 1 (Monotone feasibility in p). *For fixed q , define*

$$\Phi_q(p) := [g(\tau^{(q)}, \tau^{(p)}) \geq \rho]. \quad (\text{I})$$

If g is non-decreasing in its second argument, then $\Phi_q(p)$ is monotone in p . Hence, if a feasible p exists, the smallest feasible index

$$p_{(q)} := \min\{p : \Phi_q(p)\} \quad (\text{II})$$

is well-defined.

Proof. If $p_1 \leq p_2$ and $\Phi_q(p_1)$ holds, then by monotonicity of g in its second argument,

$$g(\tau^{(q)}, \tau^{(p_2)}) \geq g(\tau^{(q)}, \tau^{(p_1)}) \geq \rho, \quad (\text{III})$$

so $\Phi_q(p_2)$ holds. Therefore, the feasible set is a suffix in p , and the minimum exists when the set is non-empty. \square

Lemma 2 (Monotone shift in q). *Assume g is non-decreasing in its first argument. If $q' \leq q$ and both $p_{(q')}$ and $p_{(q)}$ exist, then*

$$p_{(q)} \leq p_{(q')}. \quad (\text{IV})$$

Proof. For any fixed p and $q' \leq q$,

$$g(\tau^{(q)}, \tau^{(p)}) \geq g(\tau^{(q')}, \tau^{(p)}). \quad (\text{V})$$

Hence

$$\{p : \Phi_q(p)\} \supseteq \{p : \Phi_{q'}(p)\}. \quad (\text{VI})$$

Taking minima over these sets gives $p_{(q)} \leq p_{(q')}$. \square

Lemma 3 (Loop invariant). *Let p be the pointer value at the start of the q -th outer iteration in Alg. 1. If $p_{(q)}$ exists, then*

$$p \geq p_{(q)} - 1. \quad (\text{VII})$$

Moreover, after the inner loop for this q , the algorithm stops at $p = p_{(q)} - 1$ and records $p_{(q)} = p + 1$.

Proof. Base case ($q = 1$): The algorithm sets $p \leftarrow D$, and $D \geq p_{(1)}$, so the claim holds.

Inductive step: Assume that the claim holds for q . By Lem. 1, Φ_q is monotone in p . The inner loop decreases p until $\neg\Phi_q(p)$ holds for the first time. Thus, it stops at $p = p_{(q)} - 1$, and the code sets $p_{(q)} \leftarrow p + 1$. For the next iteration, the carried pointer is $p \leftarrow p_{(q)} - 1$. By Lem. 2, $p_{(q+1)} \leq p_{(q)}$, hence

$$p = p_{(q)} - 1 \geq p_{(q+1)} - 1. \quad (\text{VIII})$$

Thus, the invariant holds for $q + 1$. \square

Proposition 1 (Correctness and time). *Assume g is non-decreasing in each argument. Then Lines 1-12 of Alg. 1 compute the frontier $\{(q, p_{(q)})\}$. If each evaluation of (f, g) on \mathcal{C} costs $\mathcal{O}(N)$ time, the total time is $\mathcal{O}(ND)$.*

Proof. By Lem. 1, each feasible $p_{(q)}$ is well-defined. By Lem. 3, at the q -th iteration the inner loop stops at $p = p_{(q)} - 1$ and records $p_{(q)} = p + 1$, which is the smallest feasible index. If no feasible p exists for some q , then $\Phi_q(D)$ is false and the guard $p_{(q)} \leq D$ excludes this q , as desired. Therefore, Lines 1-12 are correct.

For the time bound, by Lem. 2, $p_{(q)}$ is non-increasing in q . Hence, across all outer iterations, the while-guard inspects g at most D times when p is decremented and at most D additional times when the guard fails immediately at the start of an iteration, so the total number of guard evaluations of g is at most $2D$ (i.e., $\mathcal{O}(D)$). Moreover, for each recorded frontier element $(q, p_{(q)})$ (at most D in total), we use a single forward pass that computes $f(\tau^{(q)}, \tau^{(p_{(q)})})$. Each evaluation costs $\mathcal{O}(N)$. Therefore, the total time is $\mathcal{O}(ND)$. \square

Implementation note. In practice, we compute f and g simultaneously and can record their values. This merges their costs and reduces constant factors, while the asymptotic bound remains $\mathcal{O}(ND)$.

Lemma 4 (Frontier suffices). *Assume f is non-decreasing in each argument and $\mathcal{F} = \{(q, p) : g(\tau^{(q)}, \tau^{(p)}) \geq \rho\} \neq \emptyset$. For any fixed feasible q , the pair $(q, p_{(q)})$ satisfies*

$$f(\tau^{(q)}, \tau^{(p_{(q)})}) \leq f(\tau^{(q)}, \tau^{(p)}) \quad \text{for all } (q, p) \in \mathcal{F}. \quad (\text{IX})$$

Proof. By definition, $p \geq p_{(q)}$ for all feasible (q, p) . Since f is non-decreasing in its second argument,

$$f(\tau^{(q)}, \tau^{(p_{(q)})}) \leq f(\tau^{(q)}, \tau^{(p)}). \quad (\text{X})$$

\square

Table I. Performance comparisons for Qwen3-VL-MoE-30B-A3B-Instruct [26] across various expert skipping ratios.

| Method | Image Understanding | | | | | | | | Video Understanding | | | | | Avg. (%) |
|------------------------------------|---------------------|--------------|--------------|--------------|--------------|-------------|--------------|--------------|---------------------|--------------|--------------|--------------|--------------|--------------|
| | TextVQA | ChartQA | MMStar | MMBench | MMVet | MME | RealWorldQA | COCO | MVBench | EgoSchema | VMME | LVB | VMMMU | |
| $k = 8$ (Default) | 83.41 | 85.08 | 59.67 | 86.60 | 69.68 | 2500 | 66.80 | 80.37 | 64.67 | 62.45 | 54.89 | 55.42 | 47.11 | 100.00 |
| Skip 63% Experts ($\rho = 0.60$) | | | | | | | | | | | | | | |
| $k = 3$ | 80.81 | 78.12 | 66.74 | 83.33 | 68.39 | 2326 | 45.88 | 71.70 | 62.02 | 57.96 | 53.48 | 54.60 | 50.44 | 95.20 |
| NAEE [42] | 81.20 | 79.41 | 55.39 | 84.18 | 68.61 | 2348 | 59.67 | 78.09 | 61.31 | 58.32 | 51.08 | <u>55.12</u> | 48.32 | 95.61 |
| MC-MoE [22] | 82.51 | 79.37 | 56.48 | <u>86.12</u> | <u>69.37</u> | <u>2438</u> | <u>62.01</u> | <u>76.82</u> | 62.61 | 58.73 | <u>54.22</u> | 54.13 | 48.54 | <u>97.09</u> |
| DiEP [6] | <u>82.04</u> | <u>80.23</u> | <u>57.26</u> | 85.07 | 68.42 | 2405 | 60.31 | 75.41 | <u>63.15</u> | <u>59.46</u> | 53.41 | 55.08 | 48.76 | 96.80 |
| MoDES (Ours) | 81.82 | 82.48 | 58.61 | 86.17 | 69.95 | 2493 | 63.92 | 76.55 | 64.42 | 62.39 | 55.15 | 55.50 | <u>49.89</u> | 99.22 |
| Skip 75% Experts ($\rho = 0.73$) | | | | | | | | | | | | | | |
| $k = 2$ | 77.54 | 69.60 | 62.38 | 80.50 | 61.33 | 2060 | 55.56 | 82.77 | 60.70 | 53.79 | 50.67 | 54.08 | 46.00 | 92.03 |
| NAEE [42] | 78.42 | 77.28 | 54.64 | 81.34 | 65.58 | 2208 | 61.75 | <u>77.31</u> | 60.98 | 55.24 | 48.87 | <u>54.87</u> | 47.12 | 93.25 |
| MC-MoE [22] | <u>80.13</u> | 78.41 | <u>57.02</u> | <u>85.32</u> | <u>67.22</u> | <u>2286</u> | <u>61.83</u> | 74.49 | 61.65 | 57.13 | <u>52.64</u> | 54.03 | 47.49 | 95.11 |
| DiEP [6] | 79.64 | <u>78.52</u> | 56.48 | 84.91 | 67.13 | 2243 | 60.94 | 75.53 | <u>62.78</u> | <u>57.86</u> | 52.38 | 54.62 | <u>48.16</u> | <u>95.21</u> |
| MoDES (Ours) | 81.65 | 82.44 | 58.78 | 86.25 | 67.61 | 2469 | 64.71 | 75.73 | 64.45 | 62.53 | 54.81 | 55.57 | 51.22 | 99.11 |
| Skip 88% Experts ($\rho = 0.85$) | | | | | | | | | | | | | | |
| $k = 1$ | 60.71 | 52.16 | 31.63 | 54.90 | 28.07 | 1590 | 52.42 | 45.64 | 41.51 | 32.52 | 39.78 | 42.41 | 12.51 | 60.11 |
| NAEE [42] | 72.41 | 65.83 | 48.88 | 73.62 | 54.52 | 1984 | 58.62 | 60.37 | 50.24 | 49.77 | 44.48 | 45.59 | 35.57 | 80.60 |
| MC-MoE [22] | <u>74.87</u> | <u>71.43</u> | 50.74 | <u>75.42</u> | <u>61.35</u> | <u>2168</u> | 60.41 | <u>68.15</u> | 56.60 | 51.84 | <u>52.51</u> | <u>47.22</u> | <u>37.41</u> | <u>86.66</u> |
| DiEP [6] | 73.46 | 70.51 | <u>53.28</u> | 73.21 | 58.64 | 2074 | <u>63.41</u> | 62.89 | <u>57.21</u> | <u>53.61</u> | 50.78 | 46.13 | 34.79 | 85.30 |
| MoDES (Ours) | 80.97 | 78.84 | 58.18 | 85.57 | 67.75 | 2403 | 64.58 | 74.66 | 62.98 | 62.04 | 55.26 | 55.50 | 46.56 | 97.33 |

Table II. Performance comparisons for InternVL-3.5-30B-A3B-HF [57] across various expert skipping ratios.

| Method | Image Understanding | | | | | | | | Video Understanding | | | | | Avg. (%) |
|------------------------------------|---------------------|--------------|--------------|--------------|--------------|-------------|--------------|--------------|---------------------|--------------|--------------|--------------|--------------|--------------|
| | TextVQA | ChartQA | MMStar | MMBench | MMVet | MME | RealWorldQA | COCO | MVBench | EgoSchema | VMME | LVB | VMMMU | |
| $k = 8$ (Default) | 85.76 | 84.08 | 62.49 | 83.81 | 69.93 | 2312 | 64.77 | 69.30 | 68.92 | 60.49 | 58.07 | 57.64 | 45.11 | 100.00 |
| Skip 63% Experts ($\rho = 0.60$) | | | | | | | | | | | | | | |
| $k = 3$ | 82.16 | 81.38 | 60.30 | 77.94 | 68.67 | 1964 | 61.34 | 65.47 | 65.34 | 58.83 | 55.62 | 55.81 | 42.07 | 94.79 |
| NAEE [42] | 82.98 | 83.02 | 61.18 | 79.65 | 67.57 | 2054 | 61.47 | 66.05 | 66.73 | 58.46 | 56.34 | 55.74 | 42.81 | 95.86 |
| MC-MoE [22] | 84.36 | 83.22 | 61.45 | 80.89 | 68.67 | 2192 | 62.13 | 66.87 | <u>67.38</u> | <u>59.03</u> | <u>56.79</u> | 56.02 | <u>43.45</u> | <u>97.25</u> |
| DiEP [6] | 83.68 | 82.79 | <u>61.82</u> | 80.22 | 68.13 | 2084 | <u>62.56</u> | <u>67.17</u> | 66.82 | 58.74 | 56.25 | 57.84 | 43.16 | 96.82 |
| MoDES (Ours) | <u>84.27</u> | <u>83.15</u> | 62.06 | 81.46 | <u>68.41</u> | 2289 | 63.10 | 68.22 | 68.64 | 60.15 | 57.76 | <u>56.12</u> | 43.84 | 98.42 |
| Skip 75% Experts ($\rho = 0.73$) | | | | | | | | | | | | | | |
| $k = 2$ | 64.51 | 64.25 | 46.69 | 71.56 | 56.42 | 1821 | 57.29 | 58.28 | 61.42 | 53.25 | 51.06 | 48.87 | 38.63 | 83.02 |
| NAEE [42] | 75.37 | 76.18 | <u>58.82</u> | 74.53 | 61.38 | 1968 | 59.47 | 63.31 | 64.46 | 54.83 | <u>55.45</u> | 52.79 | 41.08 | 90.76 |
| MC-MoE [22] | <u>77.41</u> | 78.24 | <u>57.65</u> | 75.58 | <u>66.41</u> | <u>2037</u> | <u>60.28</u> | <u>64.24</u> | <u>65.18</u> | <u>56.14</u> | 53.65 | 53.08 | 41.74 | <u>92.30</u> |
| DiEP [6] | 76.84 | <u>79.12</u> | 58.42 | <u>76.14</u> | <u>65.27</u> | 2021 | 58.74 | <u>63.10</u> | 64.89 | 55.83 | 54.12 | <u>54.22</u> | 40.23 | 91.80 |
| MoDES (Ours) | 82.13 | 82.54 | 61.46 | 81.88 | 67.92 | 2258 | 62.48 | 67.89 | 68.83 | 60.32 | 57.54 | 55.85 | 44.16 | 97.90 |
| Skip 88% Experts ($\rho = 0.85$) | | | | | | | | | | | | | | |
| $k = 1$ | 58.49 | 46.24 | 42.27 | 51.74 | 35.05 | 1683 | 51.44 | 26.01 | 31.99 | 34.47 | 35.26 | 37.40 | 24.27 | 59.63 |
| NAEE [42] | 66.24 | 68.32 | 50.14 | 64.37 | 49.52 | 1802 | 55.23 | 50.64 | 54.78 | 50.25 | 48.69 | 47.42 | 37.27 | 78.88 |
| MC-MoE [22] | <u>70.41</u> | <u>73.49</u> | 56.14 | <u>64.38</u> | <u>65.41</u> | <u>1972</u> | <u>57.49</u> | <u>60.12</u> | <u>58.97</u> | <u>52.31</u> | <u>49.72</u> | <u>48.31</u> | <u>40.06</u> | <u>86.20</u> |
| DiEP [6] | 69.37 | 71.84 | <u>57.21</u> | 63.19 | 65.32 | 1838 | 56.38 | 55.78 | 56.26 | 51.48 | 48.94 | 47.26 | 38.18 | 83.26 |
| MoDES (Ours) | 80.58 | 82.00 | 61.20 | 81.67 | 67.80 | 2222 | 61.73 | 65.16 | 68.65 | 60.79 | 57.63 | 54.49 | 44.33 | 97.03 |

Proposition 2 (Optimality on the frontier). *Under the assumptions of Lem. 4, any optimal solution of*

$$\min_{(q,p) \in \{1,\dots,D\}^2} f(\tau^{(q)}, \tau^{(p)}) \quad \text{s.t.} \quad g(\tau^{(q)}, \tau^{(p)}) \geq \rho \quad (\text{XI})$$

lies on the frontier $\{(q, p_{(q)})\}$.

Proof. By Lem. 4, for each feasible q , the best feasible choice in p is $p_{(q)}$. Therefore, an optimal pair can be chosen from

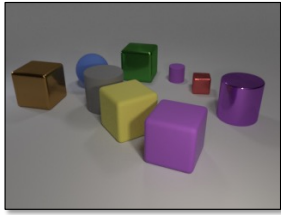
$$\{(q, p_{(q)}) : p_{(q)} \text{ exists}\}, \quad (\text{XII})$$

which is exactly the frontier. This is what Lines 13–14 minimize over, using the f -values already stored when each $(q, p_{(q)})$ was inserted into the frontier. \square

B. More Setups

Baselines. As noted in Sec. 6.1, baselines such as NAEE [42], MC-MoE [22], and DiEP [6] are not directly compatible with MoE MLLMs when $(k > 2)$. We therefore describe more about our adaptations here. For the hyperparameter $\beta^{(l)}$, we perform a genetic search under a given skipping ratio on the same dataset as our method. All remaining settings follow the original papers.

Implementation. In practice, we normalize $\alpha^{(l)}$ across layers as $\widetilde{\alpha}^{(l)} = \frac{\alpha^{(l)}}{\sum_{l'=1}^L \alpha^{(l')}}$. During inference, we compute $s_i^{(l)} = \widetilde{\alpha}^{(l)} \cdot \pi_i^{(l)}$ for a given token $\mathbf{x}^{(l)}$. Since $0 < \pi_i^{(l)} < 1$ ($i \in \mathcal{S}^{(l)}$), $s_i^{(l)} \in (0, 1)$. Thus, we choose $D = 100$ grids in $(0, 1)$ as \mathcal{B} to search for optimal thresholds. In detail, we apply a rectified sigmoid function to 100 grids falling



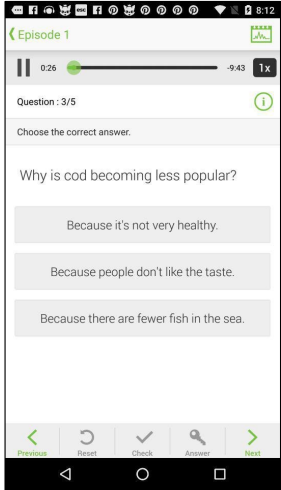
Question: Subtract all brown blocks. Subtract all large blue rubber things. Which objects are left? Use a sentence to answer the question.

Qwen3-VL-MoE-30B-A3B-Instruct : After subtracting the brown block and the large blue rubber sphere, the objects left are a green cube, a yellow cube, a purple cube, a gray cylinder, a small purple cylinder, a red cube, and a large purple cylinder.

MC-MoE: The objects left are: a gray cylinder, a yellow cube, a purple cube, and a purple cylinder.

DiEP: The objects left are a gray cylinder and a blue sphere.

MoDES (Ours): After subtracting the brown block and the large blue rubber sphere, the remaining objects are a green cube, a small purple cylinder, a small red cube, a large purple cylinder, a large yellow cube, a large purple cube, and a gray cylinder.



Question: How much time has been spent on the answer?

Qwen3-VL-MoE-30B-A3B-Instruct: Based on the image provided, we can determine the time spent on the current question by looking at the progress bar and the timer displayed.

- The progress bar shows the elapsed time and the total duration of the audio.
- The timer on the left side of the progress bar reads `0:26`.
- The timer on the right side of the progress bar shows the total duration, which is `9:43`. The negative sign indicates the time remaining, so the total duration is 9 minutes and 43 seconds.

The number `0:26` represents the amount of time that has passed since the audio started playing.

Therefore, **26 seconds** have been spent on the answer.

MC-MoE: The image provided shows a screen from a mobile application, likely a quiz or learning app, where the user is asked to choose the correct answer to a question. The question is "Why is cod becoming less popular?" with three options: "Because it's not very healthy," "Because people don't like the taste," and "Because there are fewer fish in the sea."

The time spent on the answer is not explicitly shown in the image. However, the time spent on the answer is the difference between the time spent on the current question and the time spent on the previous question. The time spent on the previous question is 0:26, and the time spent on the current question is 9:43. Therefore, the time spent on the answer is 9:43 - 0:26 = 9:17.

The time spent on the answer is **9:17**.

DiEP: The given question is "Why is cod becoming less popular?" and the options are "Because it's not very healthy", "Because people don't like the taste", and "Because there are fewer fish in the sea". The correct answer is "Because there are fewer fish in the sea". The time spent on this answer is **9:43**.

MoDES (Ours): Based on the image provided, we can determine the time spent on the answer by examining the progress bar.

- The progress bar shows the total duration of the audio clip, which is `9:43`.
- The current position of the playhead on the bar indicates how much time has passed.
- The time displayed to the left of the playhead is `0:26`.

This means that 26 seconds have elapsed since the audio began playing.

Therefore, the time spent on the answer is **26 seconds**.



Question: Based on data presented in the video, what was the approximate difference in value between the US bond market and the US stock market as of 2020?

Qwen3-VL-MoE-30B-A3B-Instruct : The video states that the bond market was valued at \$50.1 trillion and the stock market at \$40.7 trillion. The difference is calculated as \$50.1 trillion - \$40.7 trillion, which equals **\$9.4 trillion**.

MC-MoE: The US bond market value was \$50.1 trillion, while the US stock market value was \$40.7 trillion. The difference in value between the two is **\$5.1 trillion**.

DiEP: **\$50.1 trillion dollars**.

MoDES (Ours): The US bond market was valued at \$50.1 trillion, while the stock market was valued at \$40.7 trillion, resulting in a difference of approximately **\$9.4 trillion**.

Figure I. Visual understanding examples from Qwen3-VL-MoE-A3B-Instruct [26]. We employ an 88% *expert skipping ratio* for all methods, and color the text to show the **correct** or the **wrong** responses.

into $[0, 1]$ with equal intervals.

For inference speed measurement, we write efficient CUDA kernels for MoE layers. First, we implement our dual-modality thresholding method inside the router kernel, so it introduces no extra kernel launches or separate decision pass. After computing router logits and top- k , we apply a branch-free masked comparison with the modality-specific threshold and directly edit the top- k outputs: skipped routes are assigned an invalid sentinel expert id (e.g., $M+1$). During MoE dispatch/gather, sentinel en-

tries are filtered out and thus never scheduled/executed, reducing both expert compute and expert loading. The added cost is only a few element-wise operations on the small top- k list, so warp divergence/overhead is minimal and does not negate the observed wall-clock speedups. Moreover, to efficiently execute the computations for the activated experts, we employ a Grouped General Matrix Multiplication (Group GEMM) approach. Group GEMM enables the concurrent execution of all required matrix multiplications within a single, unified kernel launch. Each expert's com-



Question: Is the actress in this picture named Anne Hathaway, and which movie is this picture from?
Kimi-VL-A3B-Instruct: Yes, the actress in this picture is named Anne Hathaway, and this picture is from the movie "The Devil Wears Prada."
MC-MoE: Yes, the actress in the picture is Anne Hathaway, and this is from the movie "Rachel Getting Married."
DiEP: The actress's name is Anne Hathaway, and the movie from is Gattino.
MoDES (Ours): Yes, the actress in the picture is Anne Hathaway, and this is from the movie "The Devil Wears Prada."



Question: Please carefully observe the image and come up with a caption for the image.
Kimi-VL-A3B-Instruct: The image depicts a bunch of ripe bananas resting on a wooden surface, with a computer keyboard and a cup visible in the background, suggesting a casual or home office setting.
MC-MoE: The image features a bunch of bananas prominently displayed on a wooden table. The background is blurred, with a cup of coffee placed on the table, and a newspaper also visible. The arrangement of the objects suggests a casual, yet organized setup.
DiEP: The image depicts a stack of bananas, which are yellow and slightly green at the bottom, and a bunch of green bananas at the top. There are also some bananas in the middle of the stack, but the middle bananas are green as well.
MoDES (Ours): The image depicts a stack of bananas placed on a wooden surface, with a computer keyboard and a cup visible in the background. The bananas are yellow with some brown spots, indicating ripeness. The setting appears to be a workspace or desk area.



Question: How many pens are shown in the video?
Kimi-VL-A3B-Instruct: Three pens.
MC-MoE: Two pens.
DiEP: Two pens.
MoDES (Ours): Three pens.

Figure II. Visual understanding examples from Kimi-VL-A3B-Instruct [50]. We employ an 83% expert skipping ratio for all methods.

putation is treated as an independent sub-task within the group. The performance of this kernel is highly dependent on the workload distribution. Therefore, to achieve maximum efficiency, we perform an offline profiling step where we conduct a grid search to identify the optimal kernel tile sizes for various representative activation patterns. This ensures high computational throughput across the diverse and dynamic workloads characteristic of MoDES computation.

All performance experiments are conducted on $8 \times H200$ GPUs, and efficiency experiments are performed on a single H200 GPU.

C. More Comparison with Baselines

We provide additional results for the Qwen3-VL-MoE-30B-A3B-Instruct [26] and InternVL-3.5-30B-A3B-HF [57] in Tabs. I and II, respectively. The observations from these results align with the phenomena identified in Kimi-VL-A3B-Instruct [50]. Across different expert skipping ratios, our method consistently outperforms the baselines, with especially large gains at high skipping levels ($\geq 75\%$).

D. Visual Understanding Visualization

In this section, we present a case study comparing our proposed MoDES with previous SOTA methods [6, 22] for LLMs. As shown in Figs. I and II, MoDES consistently generates text that far outperforms the baselines.

E. Ablation for N

We apply MoDES to Kimi-VL-A3B-Instruct [50] using different numbers of data samples from GQA [25] and show the results in Tab. III. The results indicate a clear trend:

Table III. Ablation results for N .

| N | ChartQA | MME | MMBench | LVB | VMMMU |
|--|--------------|-------------|--------------|--------------|--------------|
| <i>Skip 67% Experts ($\rho = 0.65$)</i> | | | | | |
| 2048 | 88.32 | 2201 | 82.79 | 62.92 | 48.89 |
| 1024 (Ours) | <u>88.24</u> | 2204 | <u>82.73</u> | <u>62.90</u> | <u>48.78</u> |
| 512 | 87.44 | 2122 | 81.27 | 61.95 | 47.68 |
| 256 | 85.56 | 2085 | 79.68 | 60.63 | 45.11 |
| <i>Skip 83% Experts ($\rho = 0.80$)</i> | | | | | |
| 2048 | 84.84 | 2186 | 81.45 | 62.63 | <u>46.67</u> |
| 1024 (Ours) | <u>84.20</u> | <u>2162</u> | <u>81.44</u> | <u>62.60</u> | 47.11 |
| 512 | 84.12 | 2118 | 80.27 | 61.88 | 46.85 |
| 256 | 83.35 | 2016 | 77.48 | 59.84 | 43.69 |

With more calibration samples, models using expert skipping perform better. Yet the accuracy gains become smaller as the sample count grows. Moreover, doubling the samples increases both calibration and search time by $\sim 2 \times$. To balance accuracy and cost, we use 1024 samples in this paper. This choice provides most of the achievable gains while keeping computation reasonable (Sec. 6.3).

F. Ablation for D

Table IV. Ablation results for D .

| D | ChartQA | MME | MMBench | LVB | VMMMU |
|--|--------------|-------------|--------------|--------------|--------------|
| <i>Skip 67% Experts ($\rho = 0.65$)</i> | | | | | |
| 200 | <u>88.16</u> | 2219 | 82.78 | 62.94 | <u>48.76</u> |
| 100 (Ours) | 88.24 | <u>2204</u> | <u>82.73</u> | <u>62.90</u> | 48.78 |
| 50 | 87.85 | 2178 | 81.76 | 62.21 | 47.89 |
| <i>Skip 83% Experts ($\rho = 0.80$)</i> | | | | | |
| 200 | 84.78 | 2178 | 81.61 | 62.59 | 47.00 |
| 100 (Ours) | <u>84.20</u> | <u>2162</u> | <u>81.44</u> | 62.60 | <u>47.11</u> |
| 50 | 83.96 | 2143 | 80.68 | 62.47 | 47.15 |

We ablate the number of grid points D in the search space \mathcal{B} . As shown in Tab. IV, larger D brings diminishing

accuracy gains, so using a very fine grid (*e.g.*, $D > 100$) is unnecessary. The time cost also grows roughly linearly with D . Based on this trade-off, we set $D = 100$ in this work.

G. Expert Redundancy across Modalities

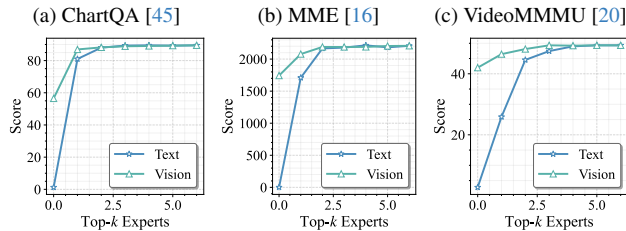


Figure III. Task performance across various numbers of top- k routed experts applied to tokens of different modalities for Kimi-VL-A3B-Instruct [50].

In this section, we analyze expert redundancy across modalities. As shown in Fig. III, reducing k for vision tokens causes task performance to drop more slowly than for text tokens. This indicates greater redundancy among experts for vision tokens, allowing more aggressive skipping than for text tokens. It also motivates modality-aware strategies for *expert skipping*.