FastMMoE: Accelerating Multimodal Large Language Models through Dynamic Expert Activation and Routing-Aware Token Pruning

Supplementary Material

A. Implementation Details and Hyperparameters Settings

For the experiments reported in the main text, the hyperparameters of **FastMMoE** are set as follows:

InternVL3.5 (MoE-based)
- **75% token retention**: $\alpha = 0.5, \gamma = 0.025, W = 5$
- **50% token retention**: $\alpha = 0.6, \gamma = 0.05, W = 5$
- **25% token retention**: $\alpha = 1.0, \gamma = 0.1, W = 5$
- **token pruning layer id**: 5,8,12

DeepSeek-VL2 (MoE-based)
- **75% token retention**: $\alpha = 0.3, \gamma = 0.05, W = 3$
- **50% token retention**: $\alpha = 0.9, \gamma = 0.05, W = 3$
- **25% token retention**: $\alpha = 0.7, \gamma = 0.2, W = 3$
- **token pruning layer id**: 2,5,8

Hyperparameter Tuning Strategy. These settings correspond to the best-performing configurations observed in the ablation studies. Across our detailed experiments, we observed a consistent pattern:
- **High Retention (e.g., 75%):** A lower α (≤ 0.5) is generally used, as attention weights constitute the primary component of the redundancy score. The merge rate γ typically adopts the theoretical upper bound (detailed in Appendix D).
- **Lower Retention (e.g., 50%, 25%):** As the retention ratio decreases, appropriately increasing both α and γ yields better performance. This validates our insight that routing similarity effectively measures token similarity, becoming increasingly critical for preserving information as pruning intensifies.
- **Model-Specific Behavior (DeepSeek-VL2):** Interestingly, for DeepSeek-VL2 at 50% and 25% retention ratios, attention weights still require a high degree of participation in the redundancy score to maintain optimal performance. We attribute this to two primary factors: (1) The presence of shared experts makes DeepSeek-VL2 structurally more similar to dense models, and attention-based visual token pruning has already been proven to be a highly effective approach for dense architectures. (2) The relatively small total number of experts in DeepSeek-VL2 (i.e., 72) may not provide sufficiently precise and effective visual token similarity information when relying heavily on routing distributions alone.

Baseline Implementation and Fair Comparison. To ensure a fair comparison of FLOPs overhead, we implemented all baseline methods (including FastV and SparseVLM) using a **multi-stage progressive pruning design**, strictly aligning the pruning layers across all methods. This multi-stage setup aligns with the core methodology and insights of PDrop [18]. FastMMoE consistently outperforms these strong multi-stage baselines.

Evaluation Setup. All evaluations are conducted using `VLMEvalKit` on **NVIDIA A800 GPUs**, with the sampling temperature fixed at 0 and sampling disabled to ensure reproducible and stable results across all benchmarks.

B. Details of Test Results

To supplement the quantitative results summarized in the main paper, this section presents the *complete benchmark results and per-configuration analyses* for both **DeepSeek-VL2** and **InternVL3.5**. These tables are provided for reproducibility and to facilitate more fine-grained comparison.

B.1. Comprehensive results on DeepSeek-VL2.

Tab. 9 reports the full benchmark performance of all pruning and acceleration methods on the **MoE-based DeepSeek-VL2**. It extends Tab. 2 in the main text by including results across six multimodal benchmarks (MMMU, SQA^I, MMBench, OCRBench, HallusionBench, AI2D). The table also differentiates between **FastMMoE*** (i.e., token pruning only without reducing activated experts) and the complete **FastMMoE** method.

From the full results we observe that, on DeepSeek-VL2, the performance gap among the three pruning methods (**FastMMoE**, **FastV**, and **SparseVLM**) is smaller than that observed on InternVL3.5. A plausible explanation is that DeepSeek-VL2's architecture includes **two shared experts** that are always activated for every token. This design is structurally closer to a dense Transformer, meaning that methods originally developed for dense models (e.g., FastV, SparseVLM) retain relatively high performance. Consequently, while **FastMMoE** still achieves the best balance between accuracy and FLOPs reduction, the relative advantage over the dense-model-oriented baselines is not as pronounced as in the InternVL3.5 experiments.

B.2. Full ablation of expert activation reduction on InternVL3.5.

Tab. 12 provides the *complete results* for the expert activation reduction strategy on **InternVL3.5-30B-A3B**, complementing the activation reduction heatmaps shown in the main text

(Section 5). Each row in Tab. 12 corresponds to a specific (K_v, l_v) configuration, where K_v is the number of experts activated per vision token and l_v is the starting layer index from which activation reduction is applied. The **Avg.**↑ column reports the average score across all six benchmarks, corresponding to the same metric definition used in the main paper. This table enables precise reproduction of the InternVL3.5 activation-reduction experiments, and it verifies that the general trends described in the main text hold consistently across all benchmarks.

B.3. Full ablation of expert activation reduction on DeepSeek-VL2.

The final long table in this section lists the complete activation reduction ablation results for **DeepSeek-VL2**, following the same structure and column definitions as Tab. 13. This facilitates direct cross-model comparison, allowing readers to observe whether the activation-reduction trends evident in InternVL3.5 experiments also manifest in DeepSeek-VL2.

All results here are obtained under the same evaluation protocol using `VLMEvalKit`, ensuring complete consistency with the benchmarks and metrics used in the main paper.

B.4. Reduce Share Experts for DeepSeek-VL2

DeepSeek-VL2 adopts a MoE architecture that contains two *shared experts* which are always activated for every token, in addition to the routed experts. These shared experts are designed to handle general-purpose features across modalities, e.g., global semantic composition and common reasoning patterns, which complement the specialized experts selected via routing. Given their architectural role, we hypothesized that reducing the number of shared experts during vision-token activation reduction could negatively impact overall performance, especially on tasks requiring fine-grained multimodal reasoning.

Tab. 8 verifies this hypothesis: when halving the number of routed experts for vision tokens (K_v from 6 to 3, $l_v = 15$) but **keeping** both shared experts active (`No` in column "Reduce ShareExperts"), average performance across benchmarks remains essentially unchanged compared to the full-expert baseline (72.83 vs. 72.77). However, if we also reduce the number of *shared experts* by half (`Yes` in "Reduce ShareExperts"), OCRBench accuracy drops sharply from 81.10 to 70.80 — a degradation of more than 10 points — while the overall average drops to 70.95. This large decline indicates that shared experts carry crucial modality-agnostic knowledge that is especially important for OCR and similar text-rich visual tasks. Given the critical nature of such tasks in real-world multimodal applications, we adopt the conservative strategy of **never reducing shared experts** in our activation-reduction pipeline for DeepSeek-VL2.

B.5. Full ablation of α and γ configurations

In the main text, Tab. 6 and Tab. 7 summarize the key results for the routing-similarity weighting parameter α and merge rate γ under several vision-token retention ratios on InternVL3.5. To provide a complete view of the parameter search space, Appendix B includes two extended tables (Tab. 14 for InternVL3.5 and Tab. 15 for DeepSeek-VL2) that enumerate the full (α, γ) combinations tested and their corresponding scores on all six benchmarks.

These extended tables complement the condensed results in the main text by showing:

- The full performance landscape across a wide range of α and γ values.
- How the optimal α and γ vary with pruning intensity (different retention ratios).
- That, consistent with the conclusions in the main paper, moderate α values and merge rates close to the theoretical upper bound provide the best trade-off between accuracy and compression.

The DeepSeek-VL2 results further confirm the general trends, while showing smaller performance fluctuations across (α, γ)—consistent with its more stable architecture that includes two permanently activated shared experts.

C. Theoretical Analysis on Expert Activation Reduction

We provide the detailed mathematical derivation to explain why reducing the number of activated experts for vision tokens causes only minor performance degradation.

C.1. Details of Vision Tokens Exhibit Magnitude Concentration

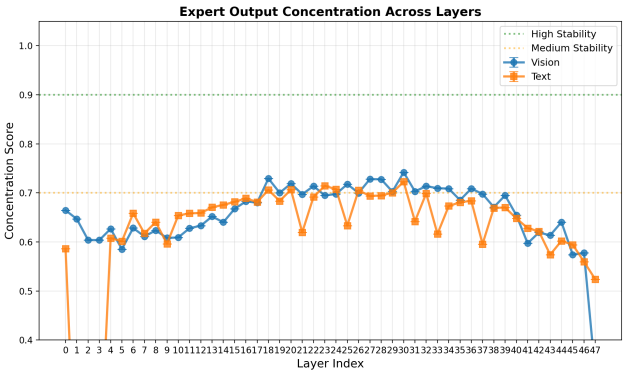


Figure 5. **Magnitude stability score V_m for InternVL3.5 across layers.** Higher V_m means tighter magnitude concentration among expert outputs.

One important empirical observation motivating our method is that the output vector magnitudes of *vision* tokens tend to be more concentrated across experts compared

Table 8. **Effect of reducing shared experts in DeepSeek-VL2 during expert activation reduction.** "Reduce ShareExperts"=Yes indicates halving the shared-expert count alongside routed experts; "No" keeps all shared experts active. Reducing shared experts causes a severe drop in OCRBench performance and a notable decrease in the overall average score, highlighting their importance for modality-general knowledge retention.

Reduce ShareExperts	K_v	l_v	MMMU	SQA^{I}	MMBench	OCRBench	HallusionBench	AI2D	Avg.↑
No	6	29	51.89	96.88	83.25	**81.40**	40.83	82.38	72.77
No	3	15	**52.11**	**96.88**	83.16	**81.10**	**41.44**	**82.29**	**72.83**
Yes	3	15	51.89	96.58	82.90	70.80	41.22	82.32	70.95

Table 9. **Peformance of all methods across different benchmarks for DeepSeek-VL2.** FastMMoE* denotes that we only apply token pruning method without reducing activated experts. The best results and second best results are indicated by **boldface** and underline, respectively. FastMMoE† denotes that we apply token pruning with reducing activated experts.

Method	MMMU	SQA^{I}	MMBench	OCRBench	HallusionBench	AI2D	Avg.↑	Drop.↓	Saving FLOPs. (%)
Baseline	**51.33**	**96.88**	**83.25**	**81.40**	**40.83**	**82.38**	**72.68**	**0**	**0**
Retain 75% vision tokens after pruning									
FastV	50.67	96.93	82.56	80.30	40.66	81.99	72.18	0.50	21.09
SparseVLM	51.11	96.53	**82.65**	78.00	40.45	81.57	71.72	0.96	
FastMMoE*	**51.56**	**96.98**	83.25	**80.80**	40.30	**82.06**	**72.49**	**0.19**	
FastMMoE† ($l_v = 10, K_v = 2$)	51.44	96.63	82.82	78.90	**40.68**	81.54	72.00	0.68	39.15
Retain 50% vision tokens after pruning									
FastV	51.00	**96.48**	82.90	79.20	**40.71**	81.19	71.91	0.77	43.07
SparseVLM	49.33	95.88	82.47	72.80	40.83	80.51	70.31	2.37	
FastMMoE*	51.11	96.33	**82.99**	**79.30**	40.55	**81.35**	**71.94**	**0.74**	
FastMMoE† ($l_v = 2, K_v = 2$)	**51.56**	95.69	82.65	75.40	37.75	80.38	70.57	2.11	61.76
Retain 25% vision tokens after pruning									
FastV	**50.56**	**95.44**	**82.13**	74.30	38.90	78.82	70.02	2.65	63.66
SparseVLM	48.11	94.74	81.44	59.00	37.33	79.02	66.61	6.07	
FastMMoE*	**50.56**	95.34	82.04	**74.70**	**40.11**	**79.15**	**70.32**	**2.36**	

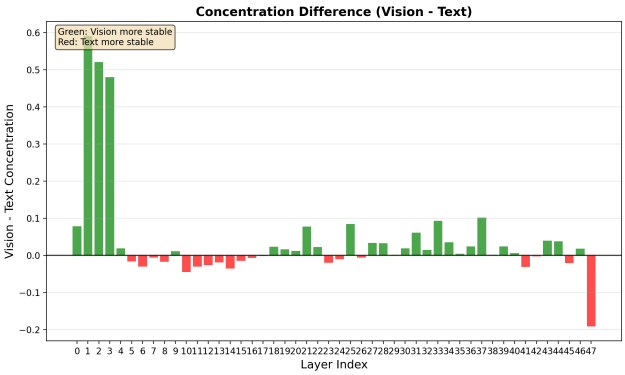


Figure 6. **Modal stability difference $V_{\text{vision}} - V_{\text{text}}$ in InternVL3.5.** Green bars: vision more stable; red bars: text more stable.

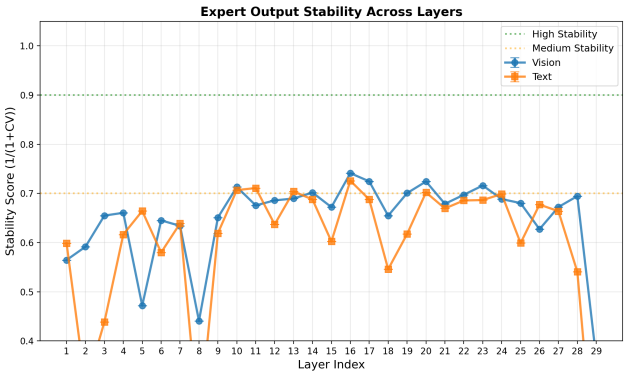


Figure 7. **Magnitude stability score V_m for DeepSeek-VL2 across layers.**

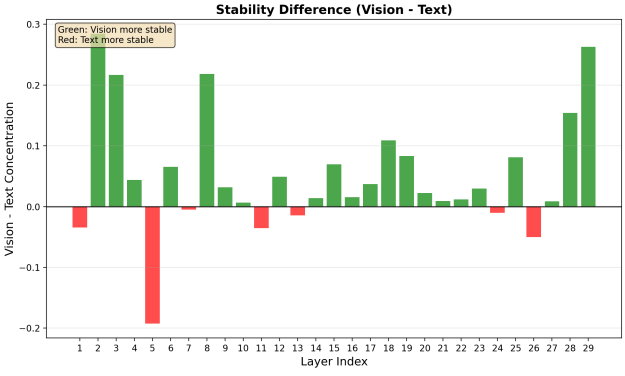


Figure 8. **Modal stability difference $V_{\text{vision}} - V_{\text{text}}$ in DeepSeek-VL2.**

to *text* tokens. This magnitude concentration means that, after activation reduction, the fused output norm changes very little; any difference between the reduced and original outputs is mainly in vector direction rather than magnitude, thereby minimizing semantic distortion while lowering computational cost.

To validate this hypothesis, we evaluate the magnitude statistics of expert outputs layer-by-layer for two representative MoE-based MLLMs: **InternVL3.5** and **DeepSeek-VL2**. For a multimodal input sequence, and for each decoder layer, we collect the intermediate outputs from all selected experts *before* the weighted fusion operation in Eq. 4. Formally, for token i of modality $m \in \{\text{Vision}, \text{Text}\}$ processed by expert e, we denote its raw output vector by

$$\mathbf{f}_{m,i}^{(e)} \in \mathbb{R}^d,$$

and compute its Euclidean norm

$$\ell_{m,i}^{(e)} = \|\mathbf{f}_{m,i}^{(e)}\|_2.$$

For each modality m in a given layer, we aggregate all such norms and compute the *coefficient of variation* (CV):

$$\text{CV}_m = \frac{\sigma_m}{\mu_m},$$

where μ_m and σ_m are the mean and standard deviation of the expert output norms for modality m. A smaller CV_m indicates that magnitudes are more tightly clustered around the mean, and hence the fused representation is more stable to changes in expert selection.

To present the results in a normalized and more interpretable way, we define the **Magnitude Stability Score**:

$$V_m = \frac{1}{1 + \text{CV}_m},$$

which maps CV values into $(0, 1]$, with higher V_m implying greater magnitude concentration. This metric is computed separately for vision and text tokens across all decoder layers.

Figs. 5 and 6 present the magnitude stability score V_m for InternVL3.5, while Figs. 7 and 8 give the same analysis for DeepSeek-VL2. In the first figure of each pair, V_m is plotted per decoder layer for both modalities, together with horizontal reference lines marking "High Stability" ($V_m \approx 0.9$) and "Medium Stability" ($V_m \approx 0.7$). In the second figure, the per-layer difference $\Delta V = V_{\text{vision}} - V_{\text{text}}$ is shown, with green bars where vision tokens are more stable and red bars where text tokens are more stable.

Across most layers in both models, vision tokens exhibit higher stability scores than text tokens, indicating more concentrated expert-output magnitudes. For InternVL3.5, large positive gaps (up to $+0.5$) appear in the early layers (Layer 2–3), while medium positive differences are common in later middle layers; a few layers towards the end show negative gaps, with text stability slightly higher. For DeepSeek-VL2, positive differences are distributed more evenly across layers, occasionally exceeding $+0.20$ in both early and late stages, but some mid-layer and output-stage layers display small negative differences (≈ -0.1 to -0.2). In both models, many layers maintain V_m values around the Medium Stability regime, supporting the observation that magnitude concentration is generally high—especially for vision tokens.

This empirical evidence verifies that vision-token expert outputs tend to have lower variance in magnitude. As a result, reweighting after pruning lower-weight experts alters the fused vector norm only marginally; the main effect of activation reduction is a slight change in output vector *direction* relative to the original, a phenomenon quantified in the subsequent theoretical analysis subsection.

Figs. 9 and 10 illustrate the distribution of expert output L2 norms across different layers for DeepSeek-VL2 and InternVL3.5, respectively. R-1 denotes the expert with the highest routing weight in the current layer, with subsequent ranks indicating decreasing weights. We observe a clear positive correlation between routing weight rank and output norm magnitude: higher-weighted experts tend to produce larger and more stable (lower CV) outputs.

Under our expert activation reduction strategy, retaining experts with the highest routing weights minimizes the loss in output norm concentration. Since these top-ranked experts dominate both the magnitude and stability of the fused representation, pruning lower-weight experts leads to only minor changes in norm distribution and, consequently, minimal performance degradation.

C.2. Setup

Let $e_i(\cdot)$ denote the transformation function of the i-th expert, and let $x \in \mathbb{R}^d$ be the hidden input vector. The routing network outputs the routing weight vector $\hat{\mathbf{a}} = G(x) \in \mathbb{R}^E$,

where E is the number of non-shared experts.

After Top-K routing selection, we choose the set $\mathcal{R} = \{r_1, r_2, \ldots, r_K\}$ of experts with the largest routing weights:

$$\hat{a}_{r_1} \geq \hat{a}_{r_2} \geq \cdots \geq \hat{a}_{r_K} > 0.$$

The weights are normalized as:

$$a_i = \frac{\hat{a}_{r_i}}{\sum_{j=1}^{K} \hat{a}_{r_j}}, \quad i = 1, \ldots, K,$$

so that $a_i > 0$ and $\sum_{i=1}^{K} a_i = 1$.

Let $p_i = e_{r_i}(x) \in \mathbb{R}^d$ denote the expert output vectors before fusion. The original MoE output is:

$$y = \sum_{i=1}^{K} a_i \, p_i. \tag{19}$$

C.3. Reduction of Activated Experts

We reduce the number of activated experts for a given token from K to m ($m < K$), keeping only the top-m weights and setting others to zero:

$$a_i' = \begin{cases} \frac{a_i}{\sum_{j=1}^{m} a_j}, & i \leq m, \\ 0, & i > m. \end{cases}$$

The reduced output is:

$$y' = \sum_{i=1}^{m} a_i' \, p_i = \sum_{i=1}^{m} \frac{a_i}{\sum_{j=1}^{m} a_j} \, p_i. \tag{20}$$

C.4. Angle Between Original and Reduced Outputs

Under the strong assumption that expert outputs are mutually orthogonal:

$$p_i^\top p_j = 0, \quad \forall i \neq j,$$

and have equal norms $\|p_i\| = 1$, the cosine of the angle θ between y and y' simplifies to:

$$\cos\theta = \frac{\sqrt{\sum_{i=1}^{m} a_i^2}}{\sqrt{\sum_{i=1}^{K} a_i^2}}, \tag{21}$$

where $a_1 \geq a_2 \geq \cdots \geq a_K$.

C.5. Lower Bound Proof

Because a_i^2 is non-negative and sorted in non-increasing order, the average of the first m terms is at least the average of all K terms:

$$\frac{1}{m} \sum_{i=1}^{m} a_i^2 \;\geq\; \frac{1}{K} \sum_{i=1}^{K} a_i^2.$$

Multiplying through by m gives:

$$\frac{\sum_{i=1}^{m} a_i^2}{\sum_{i=1}^{K} a_i^2} \;\geq\; \frac{m}{K}.$$

Taking square roots in (21) yields the tight lower bound:

$$\cos\theta \geq \sqrt{\frac{m}{K}}. \tag{22}$$

Equality occurs when all a_i are equal ($a_i = 1/K$), i.e., in the equal-weight case.

This bound is valid under the orthogonality and equal-norm assumption, and the equal-weight case corresponds to the worst possible $\cos\theta$ (largest angle). For any non-uniform distribution of a_i, the numerator $\sum_{i=1}^{m} a_i^2$ increases relative to the denominator $\sum_{i=1}^{K} a_i^2$, making the ratio larger and hence θ smaller.

C.6. Interpretation

Eq. (22) shows that, even in the worst case, reducing expert activation from K to m changes the output direction by at most $\theta_{\max} = \arccos\left(\sqrt{\frac{m}{K}}\right)$. When m/K is reasonably large, this angular deviation is small. If expert output norms are also similar, the change in output magnitude $\|y'\|$ relative to $\|y\|$ is negligible, so the primary effect of activation reduction is a slight change in direction, explaining the minimal performance drop empirically observed.

C.7. The Similarity of Experts Outputs

We measure the pairwise similarity between expert outputs for the vision modality across multiple layers in InternVL3.5. Following the definition $\text{EuclideanSim} = \frac{1}{1+\text{Dist}}$, Fig. 19 shows cosine and Euclidean similarities computed directly in the high-dimensional (2048-d) output space. In most layers, the cosine similarity between different experts is close to zero, indicating near-orthogonality and supporting the orthogonality assumption adopted in our theoretical analysis (§C).

However, the observation that expert outputs are entirely unrelated in semantics would be counter-intuitive. It is well known in high-dimensional statistics that sparse vectors tend to appear orthogonal by chance. To better reveal potential semantic correlations, we perform PCA to project expert outputs to a 32-dimensional subspace before computing similarities (Fig. 20). After dimensionality reduction, the pairwise cosine similarities increase noticeably (0.05–0.14), and Euclidean similarities also rise (up to ≈ 0.6 in some cases), suggesting that experts may share certain semantic components in lower-dimensional spaces.

This result implies that while experts appear nearly orthogonal in the original output space, their features are not completely independent in the task-relevant subspace. Consequently, when reducing the number of activated experts,

the remaining top-weight experts could partially compensate for the removed ones, resulting in an even smaller deviation angle between y and y' than predicted by the worst-case bound in Eq. (22).

C.8. Case Study on Different Models

Figures 13–15 present the routing-probability distributions for **InternVL3.5**, while Figures 16–18 show the corresponding statistics for **DeepSeek-VL2**. For each model, we visualize the mean normalized routing probabilities of Top-K experts (sorted in descending order) for both *vision* and *text* tokens across multiple layers, and plot the layer-wise sum of Top-$(K/2)$ routing probabilities.

From the InternVL3.5 results (Figs. 13 and 14), we observe that:

- Text tokens have significantly higher Top-1 and Top-2 routing probabilities than vision tokens across nearly all layers, with Top-1 probabilities often exceeding 0.27–0.30, versus 0.16–0.19 for vision tokens.
- Vision-token curves are more uniform across experts (smaller drop from Top-1 to later experts), implying weaker concentration in a few high-weight experts and a more distributed expert utilization.
- This difference is quantitatively confirmed in Fig. 15, where text tokens maintain a Top-$(K/2)$ sum well above 0.60, substantially higher than the ≈ 0.54–0.58 range for vision tokens.

According to our theoretical framework in §C, such weaker concentration for vision tokens suggests that reducing the number of activated experts (e.g., halving K) will induce even smaller angular deviation (Eq. 22) and minimal norm change, thus preserving accuracy.

For DeepSeek-VL2 (Figs. 16 and 17), the gap between text and vision tokens shows a similar but *less pronounced* pattern:

- Text tokens still show a steep drop from Top-1 to subsequent experts, with Top-1 probabilities up to 0.35–0.41 in certain layers (e.g., Layer 16, Layer 18), indicating strong expert activation concentration.
- Vision tokens' Top-1 probabilities are slightly higher relative to InternVL3.5 (around 0.17–0.19 in many layers), possibly due to the presence of two permanently activated **shared experts**, which biases the routing distribution towards high-weight slots.
- In Fig. 18, text tokens retain a Top-$(K/2)$ sum in the 0.58–0.67 range, while vision tokens are more stable around ≈ 0.52–0.53, with very low variance across layers.

The presence of shared experts in DeepSeek-VL2 reduces the relative performance gap among pruning methods (as also noted in §B), since dense-model-oriented baselines (FastV, SparseVLM) can still leverage these shared experts for stable performance. From the angle-bound perspective, the elevated and stable Top-$(K/2)$ sum for vision tokens means that activation reduction will induce bounded deviation even if $K_v \ll K$.

Conclusions. Across both models, text tokens consistently exhibit stronger expert-concentration patterns than vision tokens, reflected in higher Top-1 routing probabilities and larger Top-$(K/2)$ sums. This corroborates our *magnitude concentration* and *angular deviation bound* analyses: vision tokens' more uniform routing distributions imply that pruning lower-weight experts yields minimal representational change, especially in InternVL3.5. In DeepSeek-VL2, shared experts structurally mitigate the impact of activation reduction across modalities, explaining the smaller performance disparity among different pruning baselines observed in experiments.

C.9. Attention Stability and Representation Drift under Expert Reduction

To comprehensively address potential concerns regarding representation drift and cross-modal alignment disruption caused by reducing activated experts, we compared the internal states of the full-expert model ($K_v = 8$) with those of the reduced-expert models on the same multimodal inputs using InternVL3.5.

Attention Stability. We first evaluate whether expert activation reduction disrupts the cross-modal attention patterns, which are crucial for our routing-aware token pruning module. We compute the Spearman rank correlation of the text-to-image cross-attention weights between the baseline and the reduced models. As shown in Fig. 11, despite halving the number of active experts, the attention weights at different depths (e.g., Layer 5 and Layer 12) maintain an exceptionally high Spearman correlation (> 0.96). This strong stability confirms that the original cross-modal semantic alignment is well-preserved, thereby validating the reliability of using attention weights as a metric for token redundancy even after expert reduction.

Feature Representation Drift. Furthermore, we quantify the representation drift by measuring the cosine similarity of the aggregated visual hidden states between the full-expert baseline and reduced variants. Specifically, for a given layer l, the aggregated visual hidden state $\bar{\mathbf{h}}_v^{(l)}$ is defined as the mean over the hidden states of all N_v visual tokens in the sequence:

$$\bar{\mathbf{h}}_v^{(l)} = \frac{1}{N_v} \sum_{i=1}^{N_v} \mathbf{h}_{i,v}^{(l)}, \tag{23}$$

where $\mathbf{h}_{i,v}^{(l)}$ denotes the hidden state of the i-th visual token at layer l. We then compute the cosine similarity between $\bar{\mathbf{h}}_{v,\text{full}}^{(l)}$ and $\bar{\mathbf{h}}_{v,\text{reduced}}^{(l)}$ for each layer. As illustrated in Fig. 12,

when reducing the active experts to $K_v = 4$ (the optimal trade-off point, indicated by the green line), the visual hidden states maintain a high cosine similarity (> 0.95) with the full model across the vast majority of layers. This minimal representation drift provides a direct internal explanation for the negligible downstream performance degradation observed in Tab. 12. Conversely, extreme reduction scenarios ($K_v = 1$ or 2) lead to a sharp decline in cosine similarity, indicating significant feature collapse, which aligns with the sharp drop in benchmark accuracy for those settings.

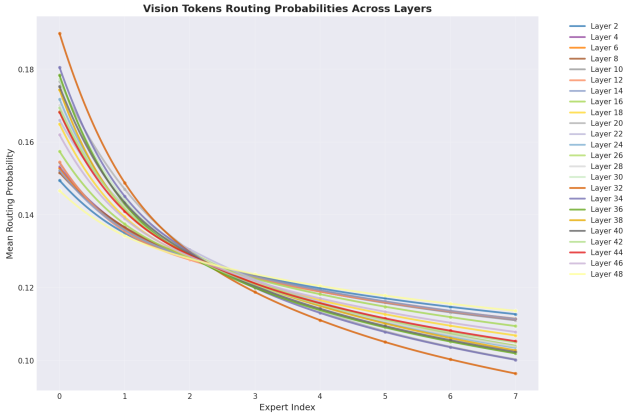


Figure 13. **InternVL3.5 Vision Tokens – Top-K Routing Probabilities Across Layers.** Mean normalized routing probabilities for the Top-K experts ($K = 8$) assigned to vision tokens, sorted in descending order per layer. The first index represents the highest-probability expert, revealing the distribution concentration towards high-weight experts across layers.

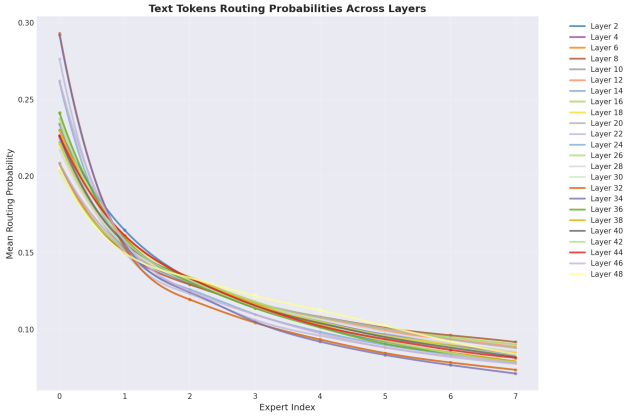


Figure 14. **InternVL3.5 Text Tokens – Top-K Routing Probabilities Across Layers.** Similar to Figure 13, but for text tokens. Text tokens display consistently higher Top-1 and Top-2 routing probabilities, indicating stronger expert activation concentration compared to vision tokens.

D. Theoretical Analysis on Merge Rate γ

This section provides the theoretical derivation of the merge rate γ used in the routing-aware token pruning module of **FastMMoE**. While the main text (§5) summarizes the empirical effects of different γ values, here we formalize the computation of its theoretical upper bound under multi-stage pruning to support the experimental configurations.

D.1. Upper Bound Derivation

The merge rate γ determines the proportion of retained vision tokens that are generated by merging high-redundancy windows, as defined in Eq. (14). In our framework, γ is upper-bounded by the extreme case in which no tokens are directly dropped; instead, all windows exceeding the redundancy threshold are merged into compressed representations until the target number of vision tokens η_l is reached. This upper bound can be derived as:

$$\hat{\gamma} = \frac{\frac{N_v^l - \eta_l}{W-1}}{\eta_l} = \frac{\frac{N_v^l}{\eta_l} - 1}{W-1} = \frac{\frac{1}{\beta} - 1}{W-1}, \tag{24}$$

where N_v^l is the number of vision tokens before pruning at layer l, W denotes the window size, and $\beta = \frac{\eta_l}{N_v^l}$ is the per-layer pruning retain ratio.

D.2. Multi-Stage Pruning Formulation

FastMMoE adopts a multi-stage pruning scheme similar to SparseVLM, applying token reduction at three designated layers. If each stage uses the same per-stage retain ratio β, the cumulative retain ratio r after three stages can be expressed as:

$$r = \beta^3, \quad \beta = r^{\frac{1}{3}}. \tag{25}$$

For target overall retain ratios of 75%, 50%, and 25%, the corresponding β values are approximately 0.91, 0.79, and 0.63. Assuming a window size $W = 5$, the maximum theoretical γ values $\hat{\gamma}$ for each setting are therefore 0.025, 0.05, and 0.15, respectively.

Table 10. **Theoretical upper bounds of the merge rate $\hat{\gamma}$.** Calculated for three target overall retain ratios (75%, 50%, 25%) under a three-stage pruning setup with equal per-stage ratio β and window size $W = 5$.

W	**Retain Ratio**	β	$\hat{\gamma}$
5	75%	0.91	0.025
	50%	0.79	0.05
	25%	0.63	0.15

D.3. Experimental Setup Based on Theoretical Bounds

In the ablation study (Table 7 in the main paper), the γ values are selected according to these theoretical bounds to ensure interpretability and consistency: for 75% retention, only $\gamma = 0.025$ is tested (its maximal feasible value); for 50% retention, $\gamma \in \{0.025, 0.05\}$; and for 25% retention, $\gamma \in \{0.025, 0.05, 0.1, 0.15\}$ to reflect the effect of stronger compression.

This theoretical analysis provides quantitative support for the chosen grid of merge rates in our experiments and helps explain the empirical observation that moderate γ values near the upper bound achieve a superior trade-off between token reduction and semantic fidelity.

E. FLOPs Analysis

In this section, we provide a complete theoretical derivation of the FLOPs savings under three configurations: (1) *token pruning only*, (2) *expert activation reduction only*, (3) *combined token pruning and activation reduction*.

E.1. Layer Index Ranges

Transformer layer index ranges for the two target architectures are:

- **DeepSeek-VL2**: 30 layers, indexed from 0 (first layer) to 29 (last layer).
- **InternVL3.5-30B-A3B**: 48 layers, indexed from 0 to 47.

All pruning layer indices in the formulas refer to these ranges.

E.2. Notation

- B: batch size (set to $B = 1$).
- L: total token length (vision + text).
- L_v: number of vision tokens before pruning.
- H: hidden size; A: attention head count; d: head dimension ($H = A \cdot d$).
- S: intermediate hidden size for dense MLP.
- S_m: intermediate hidden size for MoE experts.
- E: number of non-shared experts.
- K: baseline active experts per token.
- K_v: active experts per vision token post-reduction; $p = K_v/K$ activation ratio.
- l_v: first layer to apply activation reduction (inclusive).
- β: per-stage vision-token retention ratio after pruning.
- $C(x)$: per-layer FLOPs for a vision-token subsequence of length x in an MoE layer.
- $C_0, C_r, C_{r^2}, C_{r^3}$: shorthand for $C(L_v)$, $C(\beta L_v)$, $C(\beta^2 L_v)$, $C(\beta^3 L_v)$.
- $D_0, D_r, D_{r^2}, D_{r^3}$: FLOPs adjustments due to reduced expert activation at different pruning stages.
- $m_0, m_r, m_{r^2}, m_{r^3}$: number of layers in each stage to which activation reduction is applied.
- R: ratio of total FLOPs after a given optimization (token pruning and/or activation reduction) to the total FLOPs of the baseline model without optimization. R satisfies $0 < R \leq 1$. The **actual FLOPs saving ratio** is $1 - R$.

E.3. Baseline FLOPs

Attention:

$$\text{FLOPs}_{\text{attn}} \approx 4BL^2H + 8BLH^2. \tag{26}$$

Dense MLP:

$$\text{FLOPs}_{\text{mlp}} \approx 6BLHS. \tag{27}$$

MoE FFN:

$$\text{FLOPs}_{\text{moe}} \approx 2BLHE + 6BLHS_mK. \tag{28}$$

E.4. DeepSeek-VL2 Formulas

Pruning at layers $2, 5, 8$ with per-stage retention β, $C'(x)$ replaces K with K_v.

Token pruning only: R_{prune} is the proportion of FLOPs remaining after applying only vision-token pruning (no activation reduction); actual savings is $1 - R_{\text{prune}}$:

$$R_{\text{prune}} = \frac{6L_vHS + \sum_{t=0}^{3} C(\beta^t L_v)}{6L_vHS + 29C(L_v)}. \tag{29}$$

Activation reduction only: R_{act} is the proportion of FLOPs remaining after activation reduction from layer l_v onward, without pruning; savings is $1 - R_{\text{act}}$:

$$R_{\text{act}} = \frac{6L_vHS + l_vC(L_v) + (29 - l_v)C'(L_v)}{6L_vHS + 29C(L_v)}. \tag{30}$$

Combined: R_{comb} is the FLOPs proportion when both token pruning and activation reduction are applied; savings is $1 - R_{\text{comb}}$:

$$R_{\text{comb}} = \frac{\sum_{s\in\{0,r,r^2,r^3\}} [N_sC_s + m_s \cdot \Delta_s]}{6L_vHS + 29C(L_v)}, \tag{31}$$

where N_s is layers in stage s, $\Delta_s = C'_s - C_s$, and m_s is the number of layers in stage s affected by activation reduction.

E.5. InternVL3.5-30B-A3B Formulas

Pruning layers: $5, 8, 12$ with per-stage retention β.

Token pruning only: R_{prune} is FLOPs proportion after pruning only; savings is $1 - R_{\text{prune}}$:

$$R_{\text{prune}} = \frac{6C_0 + 3C_r + 4C_{r^2} + 35C_{r^3}}{48C_0}. \tag{32}$$

Activation reduction only: R_{act} is FLOPs proportion after activation reduction only; savings is $1 - R_{\text{act}}$:

$$R_{\text{act}} = \frac{l_v C_0 + (48 - l_v) C'(L_v)}{48 C_0}. \tag{33}$$

Combined: R_{comb} is FLOPs proportion when both strategies are applied; savings is $1 - R_{\text{comb}}$:

$$R_{\text{comb}} = \frac{\sum_{s \in \{0, r, r^2, r^3\}} [N_s C_s + m_s \cdot \Delta_s]}{48 C_0}, \tag{34}$$

where $N_s \in \{6, 3, 4, 35\}$ for $\{C_0, C_r, C_{r^2}, C_{r^3}\}$ stages, m_s is computed as defined in Notation, and $\Delta_s = C'_s - C_s$.

Interpretation

Across all configurations:

- R measures the fraction of FLOPs used by the optimized model compared to the unoptimized baseline.
- $1 - R$ is the actual FLOPs saving ratio.

For example, $R = 0.55$ means the optimized model consumes 55% of the original FLOPs, corresponding to a $1 - R = 0.45$ (i.e., 45% reduction).

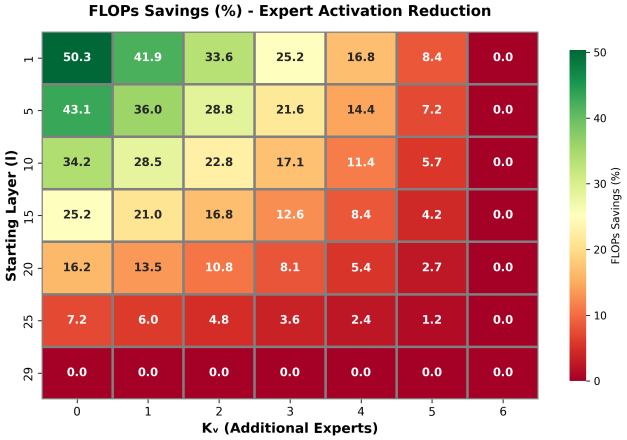


Figure 22. **FLOPs savings for DeepSeek-VL2 with expert activation reduction only.** Heatmap shows FLOPs savings (%) for different combinations of starting layer l and number of additional activated experts K_v (beyond 2 shared experts). Earlier activation reduction (smaller l) and fewer activated experts (smaller K_v) yield higher savings. Best case ($l = 1, K_v = 0$): 43.58% FLOPs reduction; worst case ($l = 29, K_v = 6$): 0% reduction.

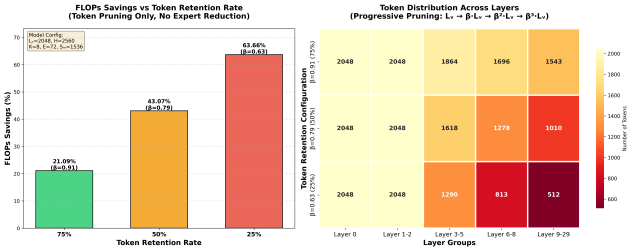


Figure 21. **FLOPs savings for DeepSeek-VL2 with token pruning only.** (*Left*) FLOPs savings at three token retention rates: $\beta = 0.91$ (75%), $\beta = 0.79$ (50%), and $\beta = 0.63$ (25%). (*Right*) Token distribution across layer groups under progressive pruning strategy ($L_v \to \beta L_v \to \beta^2 L_v \to \beta^3 L_v$). Lower retention rates yield higher FLOPs savings, with 25% retention achieving 43.74% FLOPs reduction.

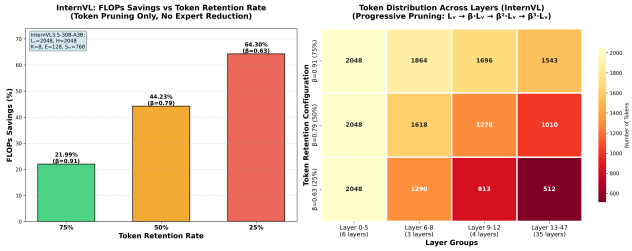


Figure 23. **FLOPs savings for InternVL3.5-30B-A3B with token pruning only.** (*Left*) FLOPs savings at three retention rates: $\beta = 0.91$ (75%), $\beta = 0.79$ (50%), and $\beta = 0.63$ (25%). (*Right*) Token distribution across four layer groups (0-5, 6-8, 9-12, 13-47) under progressive pruning. InternVL's deeper architecture (48 layers) benefits from aggressive pruning, achieving up to 64.30% FLOPs reduction at 25% retention.

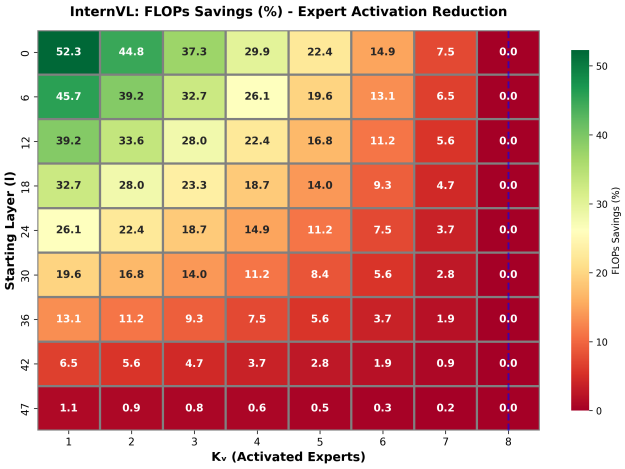


Figure 24. **FLOPs savings for InternVL3.5-30B-A3B with expert activation reduction only.** Heatmap shows FLOPs savings (%) for different starting layers l and number of activated experts $K_v \in [1, 8]$. Unlike DeepSeek-VL2, InternVL has no shared experts, allowing all experts to be freely reduced. Best case ($l = 0, K_v = 1$): 63.91% FLOPs reduction; the longer layer span (48 vs. 30) amplifies the impact of early-layer activation reduction.

E.6. Empirical FLOPs Analysis and Visualization

To validate the theoretical derivations and provide intuitive insights into the computational savings, we visualize the FLOPs reduction under different configurations for both DeepSeek-VL2 and InternVL3.5-30B-A3B. Figs. 21 to 24 present the results for token pruning only and expert activation reduction only, respectively. These visualizations reveal several key observations about the effectiveness of each strategy and how they interact with different model architectures.

Token Pruning Strategy. Figs. 21 and 23 illustrate the FLOPs savings achieved through progressive vision-token pruning at three retention rates: $\beta = 0.91$ (75%), $\beta = 0.79$ (50%), and $\beta = 0.63$ (25%). For DeepSeek-VL2, the savings range from 21.99% at 75% retention to 64.30% at 25% retention. InternVL3.5-30B-A3B exhibits a similar trend but with slightly different magnitudes due to its distinct layer structure (48 layers with pruning at layers 5, 8, 12 versus DeepSeek's 30 layers with pruning at layers 2, 5, 8). The right panels of both figures show how tokens are progressively reduced across layer groups ($L_v \to \beta L_v \to \beta^2 L_v \to \beta^3 L_v$), with the majority of layers (29 out of 30 for DeepSeek; 35 out of 48 for InternVL) operating on heavily pruned sequences ($\beta^3 L_v$). This cubic reduction in the deepest layers is the primary driver of computational savings, as the quadratic attention cost scales with the square of sequence length.

Expert Activation Reduction Strategy. Figs. 22 and 24 present heatmaps of FLOPs savings as a function of the starting layer l and the number of activated experts K_v. A critical architectural difference emerges here: DeepSeek-VL2 has 2 shared experts that are always activated, so K_v represents *additional* experts beyond the shared ones (ranging from 0 to 6 in the visualization); InternVL3.5-30B-A3B has *no* shared experts, allowing K_v to vary freely from 1 to 8. Both models exhibit a clear gradient: earlier activation reduction (smaller l) and fewer activated experts (smaller K_v) yield higher savings. For DeepSeek, applying activation reduction from layer 1 with $K_v = 0$ (only shared experts) achieves 43.58% FLOPs reduction, while delaying to layer 29 results in negligible savings. InternVL shows a similar pattern, with best-case savings of 63.91% at $l = 0, K_v = 1$. The steeper gradient in InternVL's heatmap reflects its longer layer span (48 vs. 30), which amplifies the impact of early-layer intervention.

Combined Strategy and Trade-offs. While the visualizations focus on individual strategies, our experimental results in Tab. 1 demonstrate that *combining* token pruning and expert activation reduction can yield complementary benefits. For instance, FastMMoE† with $l_v = 6, K_v = 5$ on InternVL achieves 55.02% FLOPs savings at 50% token retention while maintaining 75.63% average accuracy (only 3.53 points drop). The theoretical formulas in Sec. E account for this synergy: token pruning reduces the sequence length L_v in the attention and MoE computations, while expert activation reduction decreases the effective K in the MoE term ($6BLHS_m K$). The heatmaps reveal that *modest* activation reduction (e.g., $K_v = 4$ or 5 out of 8) from intermediate layers (e.g., $l_v = 6$ or 10) provides a favorable accuracy-efficiency trade-off, as overly aggressive reduction (e.g., $K_v = 1$ from $l = 0$) can degrade routing quality for vision tokens. This underscores the importance of tuning l_v and K_v based on the specific model architecture and downstream task requirements.

Summary. The visualizations confirm that both token pruning and expert activation reduction are effective mechanisms for reducing FLOPs in MoE-based MLLMs, with savings ranging from 20% to over 60% depending on the aggressiveness of the configuration. Token pruning is particularly impactful due to the quadratic scaling of attention, while expert activation reduction offers fine-grained control over MoE computation. The architectural differences between DeepSeek-VL2 and InternVL3.5-30B-A3B (e.g., shared experts, layer count, pruning schedules) lead to quantitatively different savings profiles, but the qualitative trends remain consistent across both models.

Table 11. **Real-world Prefill Latency on MMMU (Batch Size=8).** Evaluated on a single NVIDIA A800 GPU. FastMMoE (75% retain, 4 experts) reduces latency from 2796.5ms to 1800.3ms (**1.55× speedup**), outperforming pruning-only baselines.

Method	**Retain**	K_v **(Experts)**	**Latency (ms)**	**Speedup**
Baseline	100%	8 (Full)	2796.5	1.00×
Pruning Only	75%	8	1827.2	1.53×
FastMMoE	**75%**	**4**	**1800.3**	**1.55×**
Pruning Only	50%	8	1817.4	1.54×
Pruning Only	25%	8	1797.4	1.56×

E.7. Real-world Latency Analysis

While theoretical FLOPs reduction provides a hardware-agnostic measure of efficiency, it does not always perfectly translate to wall-clock latency improvements due to memory bandwidth constraints and kernel overheads. To evaluate the practical acceleration of **FastMMoE**, we conducted real-world prefill latency measurements on a single **NVIDIA A800 GPU** for InternVL3.5.

Table 11 reports the prefill latency on the MMMU benchmark with a batch size of 8. We observe the following key findings:

- **Significant Speedup:** Our complete FastMMoE framework (75% token retention, 4 active experts) significantly reduces the prefill latency from 2796.5 ms to 1800.3 ms, achieving a **1.55× speedup** over the full-expert baseline. This demonstrates the practical effectiveness of our method in real-world deployment scenarios.
- **Diminishing Returns at Extreme Compression:** When applying token pruning only, we observe that decreasing the retention ratio below 75% (e.g., to 50% or 25%) hits a latency plateau (around 1797 ms to 1827 ms). This indicates that at highly compressed sequence lengths, the inference becomes memory-bound, and fixed kernel overheads dominate the execution time. Consequently, the 75% retention configuration offers the optimal trade-off between latency reduction and performance preservation.
- **Potential of Expert Reduction:** Comparing FastMMoE (75% retain, 4 experts) with the Pruning-Only baseline (75% retain, 8 experts), we observe an additional latency reduction (from 1827.2 ms to 1800.3 ms). While the current lack of highly optimized, fused MoE kernels masks some of the latency gains from expert reduction, the FLOPs savings remain substantial. Future developments in fused MoE kernels are expected to further unlock the wall-clock acceleration potential of activating fewer experts.

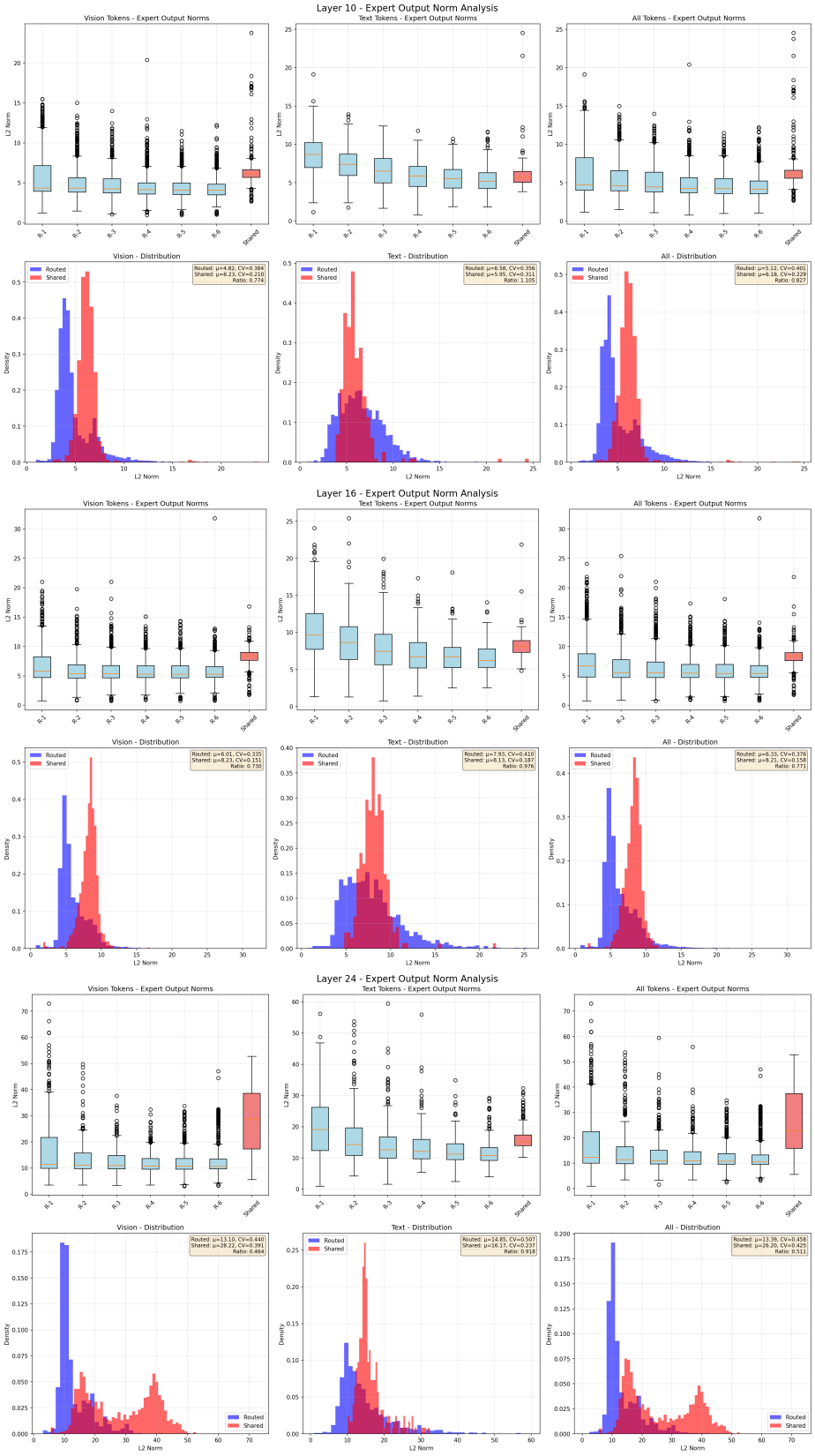


Figure 9. The experts outputs norm distribution of different modalities in DeepSeek-VL2.

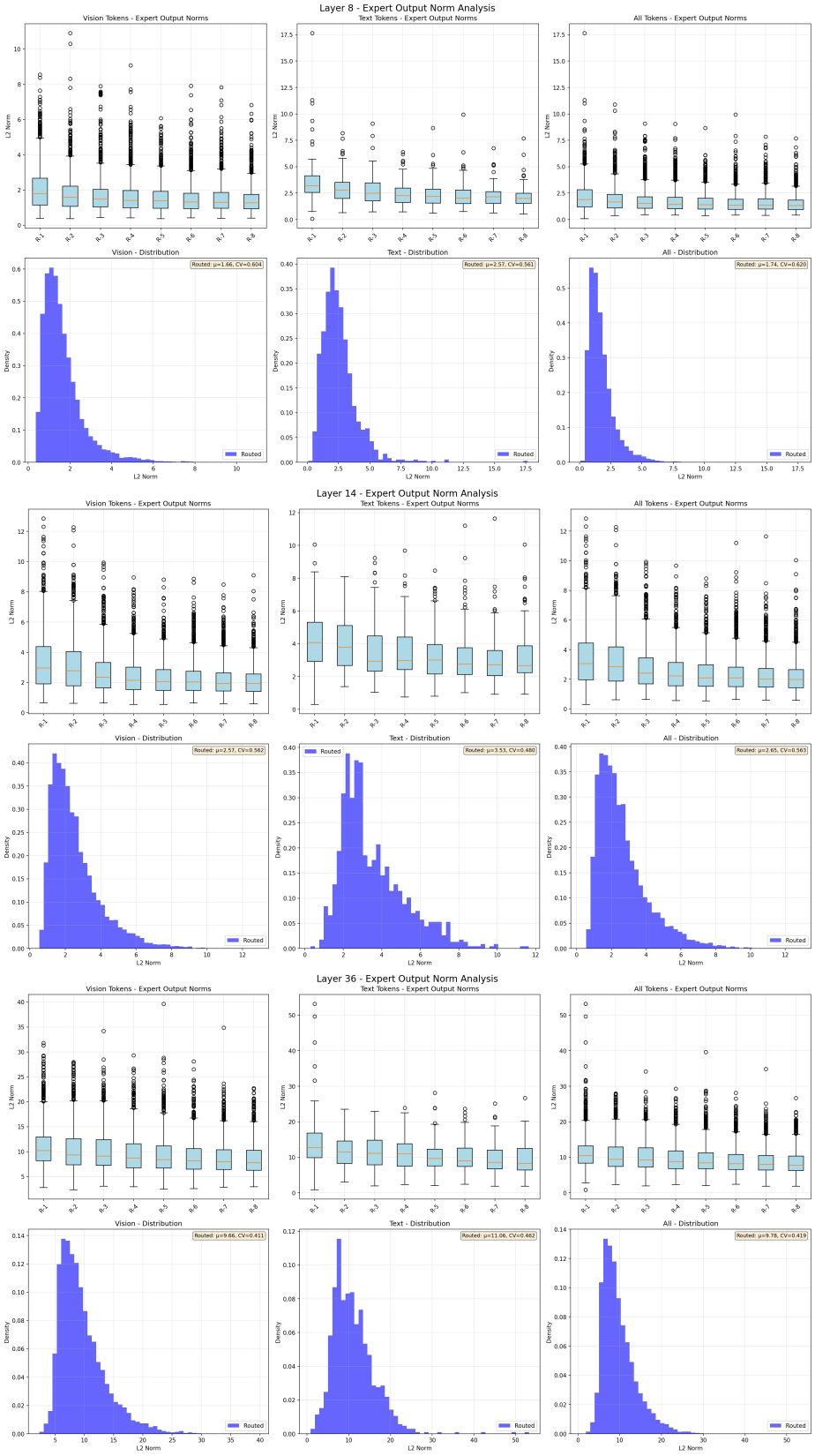
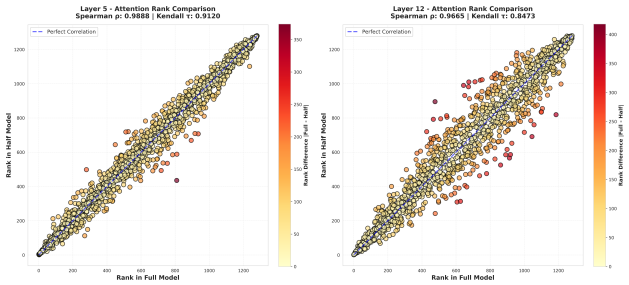


Figure 10. The experts outputs norm distribution of different modalities in InternVL3.5.



(a) Layer 5 (b) Layer 12

Figure 11. **Attention Stability under Expert Reduction.** The Spearman rank correlation of text-to-image attention weights remains highly stable (> 0.96) after reducing active experts, confirming that cross-modal alignment is preserved.

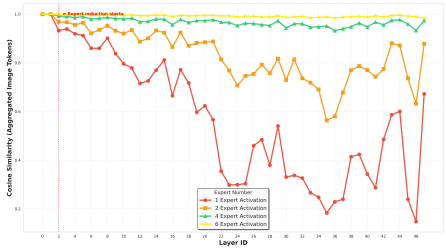


Figure 12. **Feature Cosine Similarity across Layers.** Aggregated visual hidden states maintain a high similarity (> 0.95) with the full 8-expert baseline when reduced to 4 experts (green line), demonstrating minimal representation drift.

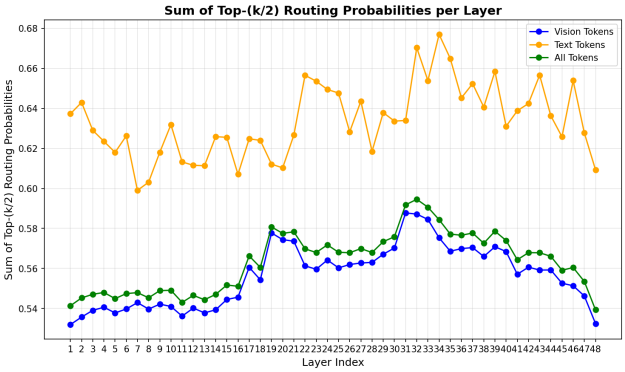


Figure 15. **InternVL3.5 – Sum of Top-$(K/2)$ Routing Probabilities Per Layer.** Layer-wise average sum of the Top-$(K/2)$=4 normalized routing probabilities for vision tokens (blue), text tokens (orange), and all tokens (green). Text tokens maintain higher concentration (> 0.60) than vision tokens (≈ 0.54–0.58) in most layers.

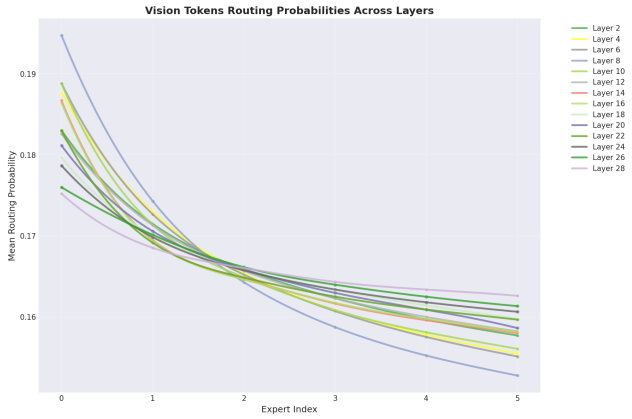


Figure 16. **DeepSeek-VL2 Vision Tokens – Top-K Routing Probabilities Across Layers.** Mean normalized routing probabilities for the Top-K=6 experts activated by vision tokens, sorted per layer. Probability drops from the highest to lowest experts are milder than in text tokens, reflecting more uniform expert usage.

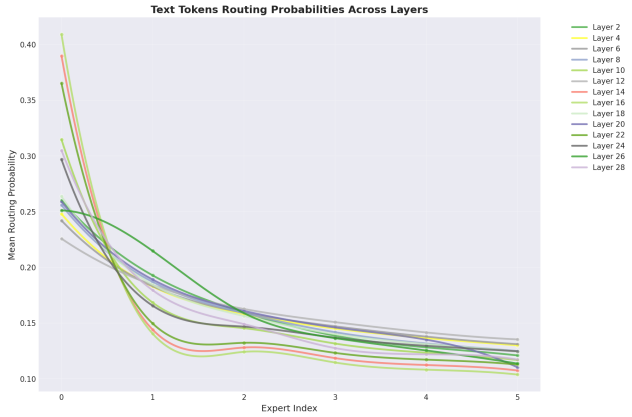


Figure 17. **DeepSeek-VL2 Text Tokens – Top-K Routing Probabilities Across Layers.** Similar to Figure 16, but for text tokens. Text tokens reach Top-1 probabilities of 0.35–0.41 in certain layers (e.g., Layer 16, Layer 18), indicating strong expert activation concentration.

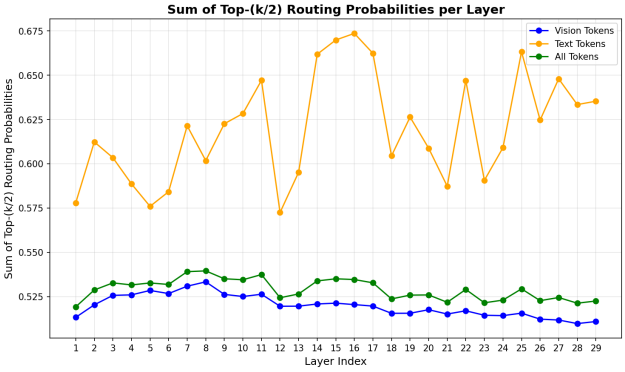


Figure 18. **DeepSeek-VL2 – Sum of Top-$(K/2)$ Routing Probabilities Per Layer.** Layer-wise average sum of the Top-$(K/2)$=3 normalized routing probabilities for vision tokens (blue), text tokens (orange), and all tokens (green). Text tokens maintain 0.58–0.67 range, whereas vision tokens remain stable around ≈ 0.52–0.53.

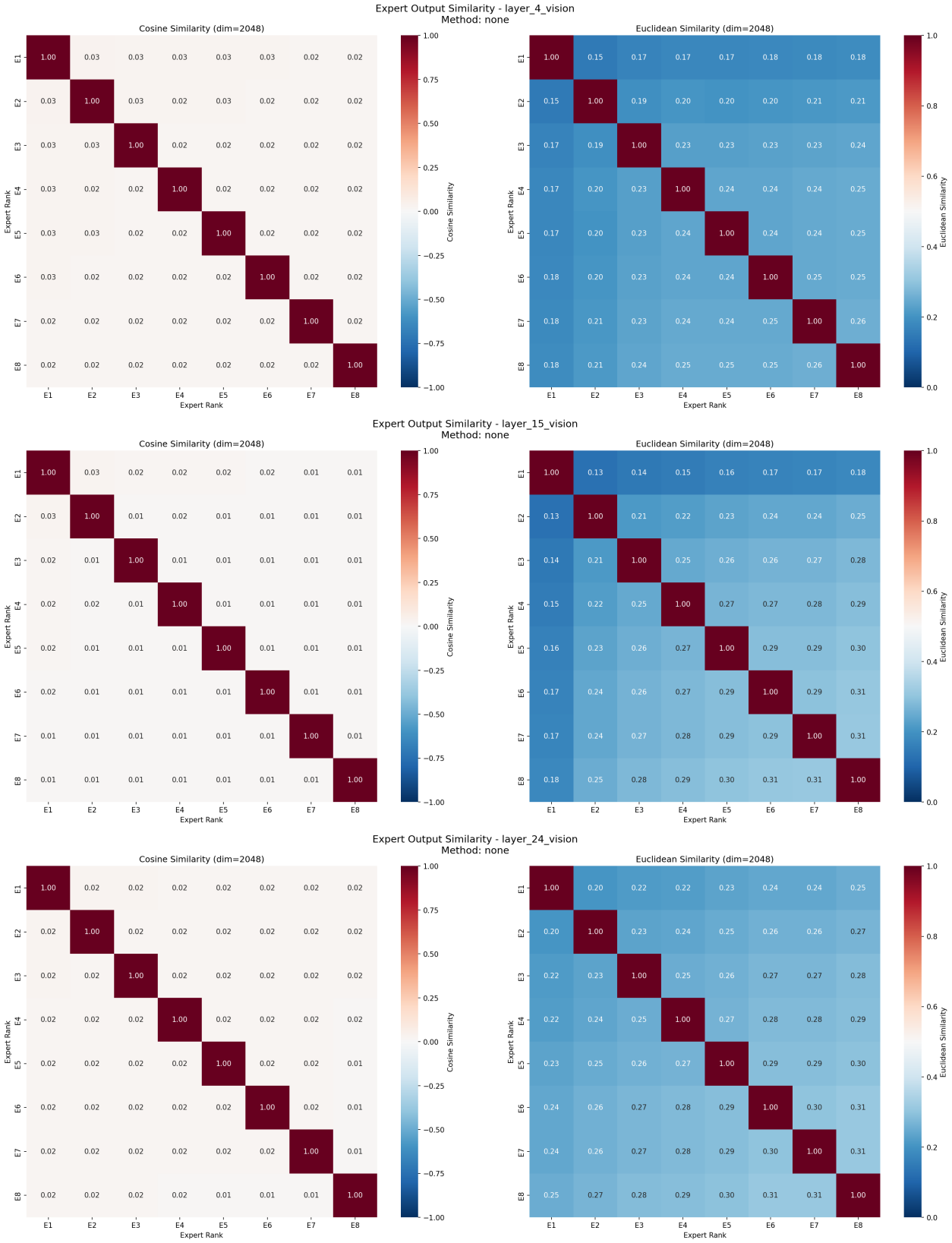


Figure 19. The cosine similarity of experts outputs across layers without PCA.

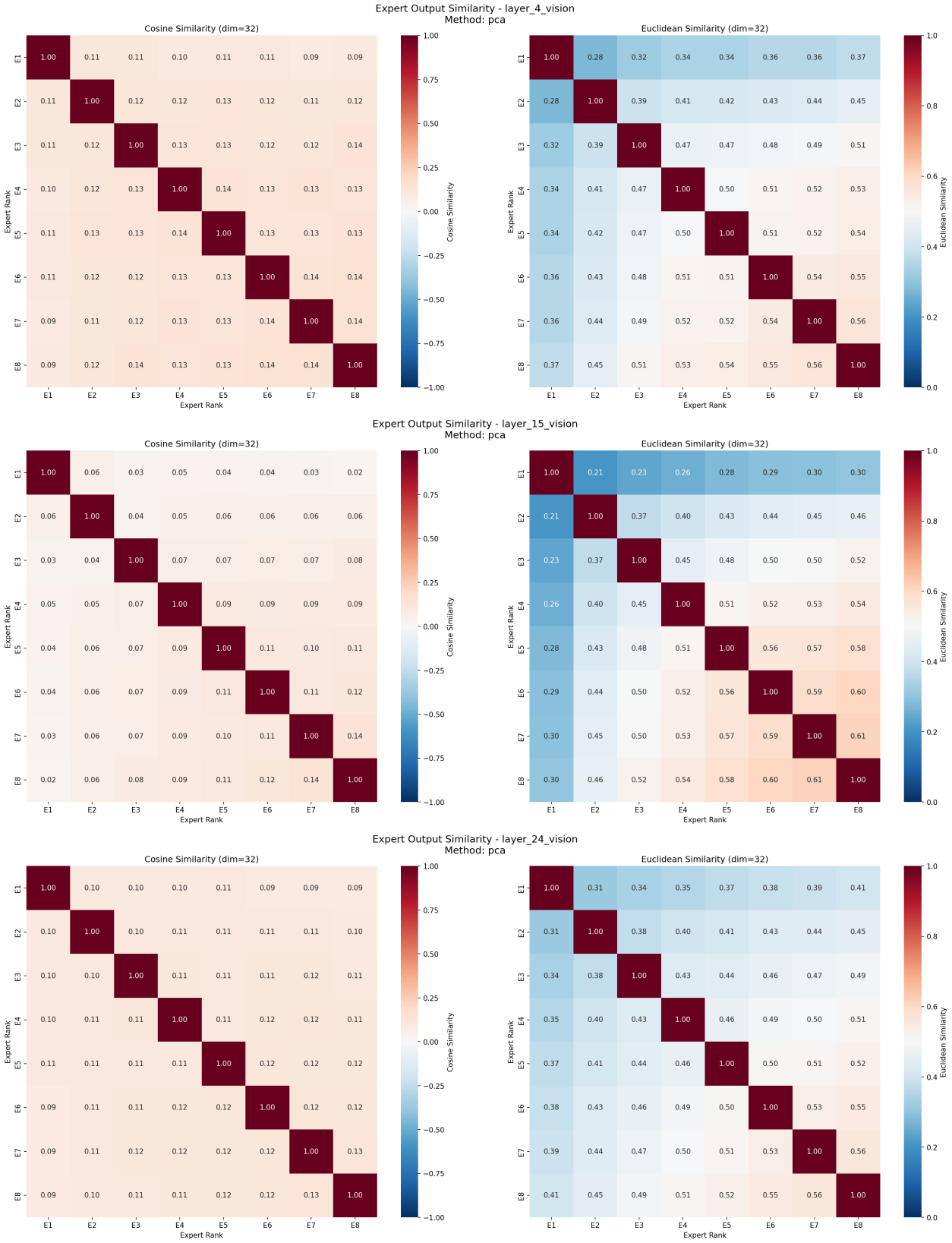


Figure 20. The cosine similarity of experts outputs across layers with PCA.

Table 12. Comparison on different strategy of reducing activation in InternVL3.5.

K_v	l_v	MMMU	SQA^{I}	MMBench	OCRBench	HallusionBench	AI2D	Avg.↑
1	2	54.22	85.37	76.63	27.70	39.91	73.12	59.49
1	5	55.22	86.32	76.55	29.70	41.38	74.06	60.54
1	10	54.11	87.70	78.52	32.10	40.50	75.65	61.43
1	15	57.00	89.89	78.78	36.90	42.09	78.01	63.78
1	20	57.56	93.60	81.27	46.30	44.96	81.61	67.55
1	25	59.78	98.56	84.88	64.50	50.20	85.36	73.88
1	30	60.33	98.91	85.91	75.90	52.28	86.66	76.67
1	35	61.67	99.01	86.34	82.80	53.41	86.95	78.36
1	40	60.89	98.96	86.43	87.00	53.43	87.05	78.96
1	45	60.89	98.96	86.43	88.00	53.46	87.18	79.15
2	2	58.78	96.38	84.71	79.30	47.56	84.36	75.18
2	5	58.33	96.88	84.28	81.30	47.02	84.33	75.36
2	10	58.33	97.37	84.88	81.20	47.21	84.91	75.65
2	15	59.33	97.87	84.71	81.40	48.57	84.91	76.13
2	20	60.11	98.31	85.22	83.70	50.69	85.75	77.30
2	25	61.11	98.71	85.91	83.90	53.65	86.24	78.25
2	30	60.44	99.01	86.08	85.80	53.48	86.88	78.62
2	35	61.00	98.96	86.43	86.60	53.66	86.92	78.93
2	40	60.56	98.96	86.34	87.60	53.56	87.11	79.02
2	45	61.00	98.96	86.34	88.20	53.25	87.08	79.14
3	2	60.44	98.86	85.82	85.70	50.02	86.04	77.82
3	5	60.56	98.66	85.65	87.00	50.84	85.82	78.09
3	10	59.78	98.86	86.43	85.90	49.67	86.24	77.81
3	15	60.67	98.91	85.48	84.80	51.27	86.27	77.90
3	20	60.22	98.81	86.25	86.10	53.02	86.66	78.51
3	25	60.11	98.86	86.43	86.80	52.24	86.66	78.52
3	30	60.78	99.01	86.25	86.40	54.01	86.76	78.87
3	35	61.11	99.01	86.25	87.40	53.86	86.98	79.10
3	40	60.56	98.96	86.43	87.80	53.56	87.24	79.09
3	45	61.00	99.01	86.43	88.20	53.14	87.11	79.15
4	2	60.11	98.91	86.34	87.20	52.41	86.46	78.57
4	5	60.11	98.96	86.51	86.90	52.35	86.24	78.51
4	10	59.78	99.01	86.34	86.90	52.34	86.82	78.53
4	15	60.78	98.91	86.25	87.60	53.02	86.46	78.84
4	20	60.89	98.91	86.25	87.50	52.95	86.85	78.89
4	25	60.67	98.91	86.25	87.40	53.26	86.82	78.88
4	30	60.33	99.06	86.60	87.20	53.69	87.21	79.01
4	35	60.44	98.96	86.34	87.40	53.66	87.08	78.98
4	40	61.00	98.96	86.43	88.00	53.77	87.11	79.21
4	45	61.11	98.96	86.43	88.40	53.14	87.18	79.20
5	2	61.11	98.86	86.43	87.50	53.06	86.72	78.95
5	5	60.56	99.06	86.60	88.20	53.07	86.82	79.05
5	10	61.00	98.91	86.60	87.80	53.57	86.88	79.13
5	15	61.78	98.96	86.17	87.60	53.77	86.79	79.18
5	20	61.33	98.86	86.60	87.50	53.54	87.21	79.17
5	25	60.67	98.86	86.34	87.70	53.84	86.88	79.05
5	30	60.33	99.01	86.34	87.50	53.59	87.21	79.00
5	35	60.78	98.96	86.34	87.90	53.66	87.11	79.12
5	40	60.89	98.96	86.51	88.20	53.25	87.11	79.15

5	45	61.00	98.96	86.43	88.50	53.03	87.18	79.18
6	2	61.33	99.01	86.43	88.00	53.32	86.95	79.17
6	5	61.44	98.91	86.25	87.90	53.66	86.95	79.19
6	10	61.56	98.91	86.43	88.30	53.78	86.88	79.31
6	15	61.22	98.86	86.00	88.00	53.17	86.95	79.03
6	20	61.44	98.96	86.43	87.70	52.90	86.88	79.05
6	25	60.89	98.96	86.60	88.00	53.08	86.95	79.08
6	30	61.00	98.96	86.43	88.20	53.46	87.01	79.18
6	35	61.22	98.96	86.43	88.60	53.56	87.18	79.32
6	40	60.78	98.96	86.51	88.40	53.36	87.18	79.20
6	45	60.89	99.01	86.34	88.50	53.03	87.08	79.14
7	2	61.67	98.91	86.51	87.90	53.70	87.05	79.29
7	5	61.33	98.91	86.08	88.10	53.35	86.85	79.10
7	10	61.33	98.91	86.17	88.00	53.53	86.95	79.15
7	15	61.67	98.86	86.25	88.20	53.66	86.85	79.25
7	20	61.33	98.96	86.25	88.10	53.60	86.76	79.17
7	25	60.56	98.96	86.25	87.90	53.12	87.24	79.00
7	30	60.67	99.01	86.51	88.00	53.27	87.11	79.10
7	35	60.78	99.01	86.43	87.90	53.56	87.18	79.14
7	40	61.11	98.96	86.51	88.50	53.25	87.18	79.25
7	45	61.00	98.96	86.43	88.50	53.14	87.05	79.18
8	48	60.67	98.96	86.43	88.60	53.14	87.14	79.16

Table 13. Comparison on different strategy of reducing activation in DeepSeek-VL2.

K_v	l_v	MMMU	SQA$^{\text{I}}$	MMBench	OCRBench	HallusionBench	AI2D	Avg.↑
0	2	47.67	90.18	79.04	62.20	32.54	78.59	65.04
0	5	50.11	91.62	80.67	71.30	34.10	80.83	68.11
0	10	51.56	96.78	82.56	80.00	41.19	82.32	72.40
0	15	51.11	96.83	83.16	81.20	41.49	82.32	72.68
0	20	51.89	96.88	83.08	81.00	41.47	82.06	72.73
0	25	51.78	96.83	83.16	81.00	41.12	82.12	72.67
1	2	49.89	93.55	81.53	70.80	35.87	80.15	68.63
1	5	50.33	94.10	81.62	75.20	37.32	80.86	69.91
1	10	51.67	96.53	82.73	78.80	39.76	81.54	71.84
1	15	51.89	96.73	83.16	79.20	41.98	82.29	72.54
1	20	51.89	96.78	83.25	81.00	41.57	82.32	72.80
1	25	51.78	96.83	83.25	80.30	41.05	82.12	72.55
2	2	51.22	96.38	82.73	78.10	38.99	81.90	71.55
2	5	51.00	96.48	83.16	79.50	39.51	81.74	71.90
2	10	51.67	96.63	82.90	80.60	40.40	81.93	72.35
2	15	52.00	96.83	83.33	80.10	41.27	82.35	72.65
2	20	51.56	96.83	83.33	80.50	41.05	82.35	72.60
2	25	51.67	96.88	83.16	81.00	40.92	82.35	72.66
3	2	52.00	96.58	83.33	79.60	40.52	82.35	72.40
3	5	51.67	96.93	82.99	80.30	40.92	82.03	72.47
3	10	52.00	97.03	83.08	80.80	40.35	81.87	72.52
3	15	52.11	96.88	83.16	81.10	41.44	82.29	72.83
3	20	51.56	96.83	83.08	80.90	41.42	82.32	72.68
3	25	51.44	96.88	83.25	81.30	40.81	82.35	72.67
4	2	51.89	96.93	83.51	80.20	41.09	81.90	72.58
4	5	52.44	96.73	83.33	80.60	41.70	82.32	72.85
4	10	51.78	96.88	83.16	80.80	41.08	82.09	72.63
4	15	51.89	96.88	83.33	81.40	40.94	82.29	72.79
4	20	51.56	96.93	83.25	81.00	41.24	82.25	72.70
4	25	51.89	96.88	83.25	81.60	40.81	82.22	72.77
5	2	52.33	96.88	83.08	80.90	41.60	81.70	72.75
5	5	51.78	96.93	83.08	81.00	40.93	82.35	72.68
5	10	52.56	96.88	83.16	80.90	41.06	82.12	72.78
5	15	52.11	96.88	83.08	81.50	41.14	82.06	72.79
5	20	51.56	96.88	83.25	81.40	40.90	82.22	72.70
5	25	51.67	96.83	83.25	81.60	40.97	82.29	72.77
6	29	51.89	96.88	83.25	81.40	40.83	82.38	72.77

Table 14. **Comprehensive ablation study of routing-similarity weighting parameter α and merge rate γ on InternVL3.5.** We report results across six benchmarks under three different vision-token retain ratios (75%, 50%, and 25%). α controls the balance between routing-probability similarity and cross-modal attention in redundancy scoring, while γ controls the proportion of retained tokens produced through merging high-redundancy windows. The table enumerates all (α, γ) combinations tested, showing that optimal values of α and γ vary with the pruning intensity: moderate α values and merge rates close to their theoretical upper bound yield the best trade-off between accuracy and compression.

Retain Ratio	α	γ	MMMU	SQA$^{\text{I}}$	MMBench	OCRBench	HallusionBench	AI2D	Avg.↑
	1	0.025	60.44	98.96	86.25	80.00	52.21	86.33	77.37
	0.9	0.025	60.78	99.01	86.25	79.30	51.40	86.43	77.20
75%									

	0.8	0.025	60.89	98.91	86.51	79.90	51.45	86.30	77.33
	0.7	0.025	60.78	98.91	86.60	80.40	51.26	86.08	77.34
	0.6	0.025	60.33	98.86	86.17	79.70	52.29	86.20	77.26
	0.5	0.025	60.67	98.91	86.60	80.70	52.21	86.27	77.56
	0.4	0.025	**60.78**	99.01	86.25	80.50	52.46	86.20	77.53
	0.3	0.025	60.78	98.91	86.25	80.50	51.73	86.08	77.37
	0.2	0.025	60.78	99.01	86.00	80.70	51.95	86.20	77.44
	0.1	0.025	60.78	98.86	86.00	79.50	52.26	86.08	77.24
	0	0.025	60.78	98.96	86.17	79.70	51.41	86.08	77.18
50%	1	0.05	59.89	98.56	85.57	76.80	50.92	84.91	76.11
	1	0.025	60.56	97.87	85.91	75.70	50.62	85.07	75.95
	0.9	0.05	**60.56**	98.56	85.22	75.70	52.58	84.78	76.23
	0.9	0.025	60.44	98.22	85.91	74.30	50.56	84.13	75.59
	0.8	0.05	60.89	98.41	85.31	75.70	51.23	84.94	76.08
	0.8	0.025	60.44	98.17	85.14	75.40	49.98	84.07	75.53
	0.7	0.05	60.00	98.26	85.31	75.30	53.25	84.49	76.10
	0.7	0.025	59.89	98.17	85.31	74.30	49.73	84.29	75.28
	0.6	0.05	61.00	98.36	85.22	76.50	52.54	84.49	76.35
	0.6	0.025	60.44	98.22	85.05	74.10	48.57	84.26	75.11
	0.5	0.05	59.56	98.07	85.65	76.00	50.00	84.59	75.64
	0.5	0.025	60.56	98.12	84.97	72.70	48.04	83.48	74.64
	0.4	0.05	59.89	98.22	85.40	76.00	50.52	84.72	75.79
	0.4	0.025	59.56	98.26	84.97	74.10	48.61	83.48	74.83
	0.3	0.05	60.22	98.36	85.40	75.70	49.66	84.81	75.69
	0.3	0.025	60.89	98.12	84.88	74.10	47.03	83.42	74.74
	0.2	0.05	60.89	98.07	85.74	74.80	49.72	84.13	75.56
	0.2	0.025	59.89	98.02	84.79	73.70	48.66	83.81	74.81
	0.1	0.05	60.22	97.97	85.22	75.80	50.56	84.07	75.64
	0.1	0.025	60.89	98.02	85.14	74.10	48.61	83.39	75.02
	0	0.05	60.22	98.22	85.48	75.30	49.72	83.65	75.43
	0	0.025	60.00	97.87	85.31	73.50	49.00	83.03	74.78
25%	1	0.15	59.33	96.18	83.76	69.20	46.77	80.99	72.71
	1	0.1	60.78	96.98	83.93	66.80	50.52	82.12	73.52
	1	0.05	59.00	97.03	82.73	62.30	44.84	80.44	71.06
	1	0.025	59.44	96.18	82.56	60.00	45.59	78.72	70.42
	0.9	0.15	59.22	96.93	84.28	66.90	46.18	81.57	72.51
	0.9	0.1	60.00	96.63	83.33	64.20	50.33	81.02	72.59
	0.9	0.05	58.67	96.58	82.30	61.50	47.48	78.82	70.89
	0.9	0.025	59.44	96.13	81.79	56.80	44.05	78.11	69.39
	0.8	0.15	59.67	96.73	83.85	66.00	47.15	81.06	72.41
	0.8	0.1	59.44	96.28	82.99	63.80	49.91	81.09	72.25
	0.8	0.05	59.56	96.13	82.90	58.60	46.06	78.37	70.27
	0.8	0.025	57.89	95.54	82.73	55.80	46.02	78.04	69.34
	0.7	0.15	60.00	96.68	83.59	65.60	47.72	80.99	72.43
	0.7	0.1	60.33	96.63	83.51	63.70	47.44	80.31	71.99
	0.7	0.05	58.44	96.13	83.08	59.40	45.24	78.21	70.08
	0.7	0.025	58.78	95.64	82.90	56.30	44.85	77.95	69.40
	0.6	0.15	60.00	96.73	83.33	64.90	47.20	80.96	72.19
	0.6	0.1	60.00	95.98	83.68	62.50	46.97	80.28	71.57
	0.6	0.05	58.44	95.98	82.73	58.20	44.74	77.91	69.67
	0.6	0.025	58.56	95.88	82.13	55.30	45.63	78.01	69.25

0.5	0.15	58.33	96.23	82.82	64.50	46.10	81.41	71.57
0.5	0.1	58.78	95.74	83.51	61.30	47.27	79.92	71.09
0.5	0.05	58.33	95.49	82.82	57.80	45.71	78.53	69.78
0.5	0.025	58.78	95.79	82.47	55.70	44.73	77.30	69.13
0.4	0.15	59.67	96.28	83.33	65.10	46.80	80.28	71.91
0.4	0.1	59.78	95.88	82.99	60.80	48.75	79.37	71.26
0.4	0.05	58.11	96.03	83.16	56.90	45.92	78.40	69.75
0.4	0.025	57.67	95.64	82.39	56.10	44.53	77.88	69.03
0.3	0.15	59.89	96.33	83.16	64.90	47.39	79.89	71.93
0.3	0.1	59.67	95.98	83.16	61.70	47.57	79.08	71.19
0.3	0.05	58.33	96.13	82.90	57.00	45.99	78.17	69.76
0.3	0.025	58.56	95.54	82.30	56.20	44.17	77.49	69.04
0.2	0.15	58.56	96.48	82.13	63.10	46.85	80.18	71.22
0.2	0.1	59.22	96.08	83.59	60.80	47.23	79.83	71.13
0.2	0.05	58.22	95.98	83.08	56.90	45.12	78.01	69.55
0.2	0.025	59.00	95.49	82.22	56.30	45.82	77.66	69.41
0.1	0.15	59.33	96.33	82.82	63.00	47.37	80.34	71.53
0.1	0.1	59.33	96.13	82.99	60.90	46.12	79.83	70.88
0.1	0.05	57.89	95.79	82.56	56.00	44.86	78.47	69.26
0.1	0.025	57.89	95.49	81.96	55.70	45.49	77.49	69.00
0	0.15	59.00	96.23	82.47	62.90	46.81	79.73	71.19
0	0.1	59.56	96.38	82.99	60.00	44.81	78.82	70.43
0	0.05	58.00	96.03	83.08	57.60	45.36	78.34	69.73
0	0.025	57.78	95.54	82.04	56.60	46.38	78.24	69.43

Table 15. **Comprehensive ablation study of routing-similarity weighting parameter α and merge rate γ on DeepSeek-VL2.** Results are reported for three retain ratios (75%, 50%, and 25%) across six benchmarks. α and γ are defined as in Table 14. Compared to InternVL3.5, DeepSeek-VL2 shows smaller performance fluctuations across (α, γ) settings, partly due to its two always-activated shared experts, which stabilize performance under pruning. Nonetheless, the trends remain consistent: balanced α values and moderate γ provide the best retention-performance trade-off.

Retain Ratio	α	γ	**MMMU**	**SQA$^{\text{I}}$**	**MMBench**	**OCRBench**	**HallusionBench**	**AI2D**	**Avg.↑**
75%	1	0.05	51.44	97.03	82.47	80.00	41.07	81.99	72.33
	0.9	0.05	51.11	96.93	82.65	79.50	40.76	81.90	72.14
	0.8	0.05	51.56	96.78	82.99	79.90	40.89	81.99	72.35
	0.7	0.05	51.67	96.83	82.90	80.40	41.08	81.74	72.44
	0.6	0.05	51.44	96.83	82.73	80.60	40.89	81.67	72.36
	0.5	0.05	52.00	97.03	82.90	80.70	40.69	81.90	72.54
	0.4	0.05	51.11	96.98	83.08	80.10	41.23	81.77	72.38
	0.3	0.05	51.56	96.98	83.25	80.80	40.30	82.06	72.49
50%	1	0.1	50.89	96.38	82.82	77.80	39.80	81.70	71.56
	1	0.05	50.11	96.53	82.82	79.70	40.68	81.22	71.84
	0.9	0.1	50.56	96.38	83.08	79.40	40.06	81.25	71.79
	0.9	0.05	51.11	96.33	82.99	79.30	40.55	81.35	71.94
	0.8	0.1	50.67	96.38	82.99	79.30	40.20	81.25	71.80
	0.8	0.05	50.22	96.48	83.08	79.20	40.07	81.15	71.70
	0.7	0.1	50.11	96.18	82.65	78.80	39.32	81.06	71.35
	0.7	0.05	51.00	96.18	82.73	79.20	40.23	81.09	71.74
	0.6	0.1	49.89	96.63	82.73	79.20	39.12	81.31	71.48
	0.6	0.05	50.44	96.43	82.99	79.80	39.94	80.83	71.74

	0.5	0.1	50.00	96.43	82.90	79.00	40.79	81.09	71.70
	0.5	0.05	50.78	96.48	83.16	80.00	40.13	80.99	71.92
	0.4	0.1	50.22	96.43	82.99	79.00	39.56	81.06	71.54
	0.4	0.05	50.67	96.63	82.82	79.90	39.20	81.19	71.73
	0.3	0.1	50.78	96.38	83.08	79.10	39.69	80.51	71.59
	0.3	0.05	50.67	96.38	82.82	79.70	39.82	80.99	71.73
	1	0.25	50.00	95.19	80.33	70.20	39.23	79.27	69.04
	1	0.2	49.89	95.24	81.53	71.40	38.96	79.50	69.42
	1	0.15	51.00	95.74	82.13	72.70	38.61	79.18	69.89
	1	0.1	50.44	95.09	81.62	73.10	39.35	79.08	69.78
	1	0.05	50.78	95.34	81.96	74.50	39.18	79.31	70.18
	0.9	0.25	50.00	95.09	80.93	73.50	38.36	78.95	69.47
	0.9	0.2	50.33	95.19	81.96	73.90	39.14	79.08	69.93
	0.9	0.15	50.00	95.29	81.53	74.50	39.02	78.98	69.89
	0.9	0.1	50.33	95.59	81.96	74.20	39.02	78.85	69.99
	0.9	0.05	50.33	95.59	82.22	74.90	38.84	78.95	70.14
	0.8	0.25	51.00	95.24	80.84	73.00	38.83	78.85	69.63
	0.8	0.2	51.11	95.29	82.04	74.10	38.36	79.08	70.00
	0.8	0.15	49.78	95.49	82.04	74.20	38.89	78.95	69.89
	0.8	0.1	50.67	95.79	81.79	73.80	39.71	79.40	70.19
	0.8	0.05	50.56	95.49	82.04	73.70	39.06	78.82	69.95
	0.7	0.25	50.44	95.14	81.19	73.60	38.61	78.95	69.66
	0.7	0.2	50.56	95.34	82.04	74.70	40.11	79.15	70.32
	0.7	0.15	49.89	95.74	82.13	74.10	38.92	79.21	70.00
	0.7	0.1	49.78	95.44	81.79	74.10	39.09	79.18	69.90
25%	0.7	0.05	49.56	95.49	82.04	74.30	38.76	78.98	69.86
	0.6	0.25	49.89	95.14	81.27	73.70	38.51	78.76	69.54
	0.6	0.2	49.89	95.49	81.44	74.10	39.30	78.76	69.83
	0.6	0.15	50.44	95.29	81.53	74.40	39.48	78.98	70.02
	0.6	0.1	49.78	95.24	82.13	73.60	39.19	79.27	69.87
	0.5	0.25	50.33	95.34	81.79	73.10	38.61	78.82	69.67
	0.5	0.2	50.11	95.69	81.36	74.90	39.46	79.15	70.11
	0.5	0.15	49.78	95.59	82.39	74.10	39.29	78.98	70.02
	0.5	0.1	50.11	95.39	82.13	73.80	38.66	79.05	69.86
	0.5	0.05	50.78	95.44	81.79	73.70	39.16	78.85	69.95
	0.4	0.25	49.78	95.19	81.10	73.20	38.63	79.08	69.50
	0.4	0.2	50.67	95.39	81.79	74.80	38.70	78.95	70.05
	0.4	0.15	50.22	95.24	82.39	74.20	38.44	79.18	69.95
	0.4	0.1	49.89	95.39	82.13	73.60	38.87	79.15	69.84
	0.4	0.05	50.78	95.79	81.70	74.30	39.36	79.08	70.17
	0.3	0.25	49.89	95.19	81.53	74.40	38.41	78.82	69.71
	0.3	0.2	50.22	95.44	81.70	74.50	38.90	79.08	69.97
	0.3	0.15	50.00	95.44	81.87	74.30	38.81	79.08	69.92
	0.3	0.1	49.78	95.39	82.04	74.10	38.94	79.02	69.88
	0.3	0.05	50.00	95.54	82.04	74.10	38.96	79.05	69.95