Supplementary materials for *AnyExperts: On-Demand Expert Allocation for Multimodal Language Models with Mixture of Experts*

Yuting Gao[1*] Lan Wang[1*] Hengyuan Zhao[12*] Linjiang Huang[2] Si Liu[2] Qingpei Guo[1]
[1]Ant Group [2]Beijing University of Aeronautics and Astronautics
yutinggao.sh@gmail.com

In this appendix, we provide supplementary results and analyses omitted from the main paper due to page constraints. Specifically, Section 1 validates that a 10% reduction in average activated experts (from 8.0 to 7.2) preserves accuracy across all modalities, as noted in Table 1; Section 2 includes the sensitivity analysis for audio understanding, omitted from Figure 3 due to its robustness to expert budget changes; Section 3 provides per-benchmark performance details underlying the averaged scores reported experiment part; Finally, Section 4 provides more visualization examples.

1. Comp-Acc Trade-off

In this section, we present the full performance of AnyExperts on multimodal benchmarks under two expert activation budgets: the default Avg-$K = 8.0$ and a reduced Avg-$K = 7.2$ (a 10% decrease). As shown in Table 2, the model maintains consistently strong results across all modalities (general vision-language, OCR, NLP, speech, and video), when using fewer activated experts. In most cases, performance remains nearly unchanged, and in some benchmarks (*e.g.*, MMVet, OvOBench), it even slightly improves. The average scores across each modality category show negligible degradation (or minor gains), confirming that AnyExperts achieves high efficiency without sacrificing accuracy. This validates our claim in the main paper that a modest reduction in expert usage preserves overall capability, making the model more computationally economical while retaining robustness.

2. Audio Robustness

In this section, we present a detailed ablation on the number of activated experts for audio understanding, which was omitted from Figure 3 due to its robustness to expert budget variations. As shown in Table 1, the Word Error Rate (WER) remains largely stable across different activation budgets, both for our dynamic AnyExperts (with average *Avg-K*) and static top-K baselines. This insensitivity suggests that the audio modality in AnyExperts is particularly robust to routing granularity, in contrast to other modalities that exhibit stronger dependence on expert allocation.

Table 1. Ablation on the number of activated experts (Audio-only results). Gray rows: our dynamic AnyExperts model with average activation count (*Avg-K*). White rows: static top-K routing baseline with fixed K per token.

Experts / Token	Audio (WER ↓)
Dynamic (AnyExperts, Ours)	
Avg-K=4.8	2.99
Avg-K=5.6	2.92
Avg-K=6.4	2.88
Avg-K=7.2	2.82
Avg-K=8.0	2.82
Avg-K=8.8	2.78
Avg-K=9.6	2.78
Static (Top-K Routing Baseline)	
K=4	2.98
K=6	2.82
K=8	2.84
K=10	2.78

3. Per-Benchmark Results

In this section, we provide per-benchmark performance details underlying the averaged scores reported in the main ablation study. The Table 3 includes variants that remove HSM or IAR, as well as sensitivity analyses over hyperparameters α, $\beta_{\max}$, and alternative MLP architectures (WildMLP and DeepMLP).

4. More Visualization

We performed further visualization analyses on multimodal data across diverse task types. For the general task in Figure 1, we evaluated the model's performance in target spatial localization, color, position, category recognition, and common sense reasoning. Specifically, in the sample of the first column, the model successfully localized two

*The first three authors contributed equally.

Table 2. Performance of AnyExperts on multimodal benchmarks with default expert activation budget (Avg-K=8.0).

General ↑			OCR ↑			NLP ↑			Speech ↓			Video ↑		
Bench	**8.0**	**7.2**	**Bench**	**8.0**	**7.2**	**Bench**	**8.0**	**7.2**	**Bench**	**8.0**	**7.2**	**Bench**	**8.0**	**7.2**
MMBench	79.73	79.73	ChartQA	83.04	82.64	ARC-C	90.17	91.19	Aishell1	1.70	1.68	MVBench	66.40	66.00
AI2D	81.19	81.19	TextVQA	73.42	73.08	GSM8k	84.76	84.61	Aishell2-ios	2.78	2.78	VideoMME	60.59	60.30
MMMU	49.11	48.67	OCRBench	85.10	85.00	GPQA	41.92	38.87	Aishell2-android	2.76	2.78	OvOBench	44.83	46.22
MMVet	66.15	66.97							Librispeech-other	3.01	3.00			
MathVista	66.70	66.63							Librispeech-clean	1.38	1.38			
MMStar	60.48	60.90												
Average	67.22	67.35	Average	80.52	80.24	Average	72.28	71.56	Average	2.32	2.32	Average	57.27	57.51

Table 3. Ablation study of AnyExperts. Column groups correspond to: Ours (baseline model), w/o HSM, w/o IAR, α-ablation ($\alpha = 0.005, 0.05, 0.1$), β_{max}-ablation, and MLP variants, WMLP denote WildMLP and DMLP represent DeepMLP. ↑: higher is better; ↓: lower is better.

Modality	**Benchmark**	**Ours**	**w/o HSM**	**w/o IAR**	$\alpha = 0.005$	$\alpha = 0.05$	$\alpha = 0.1$	$\beta_{\max} = 0.1$	$\beta_{\max} = 0.25$	**WMLP**	**DMLP**
General (↑)	MMBench	66.84	61.17	60.13	61.08	66.15	65.12	60.91	62.63	65.89	61.00
	AI2D	67.88	65.71	66.13	65.64	68.17	68.85	67.58	66.39	68.20	67.23
	MMMU	41.89	42.11	40.78	41.44	45.22	43.22	44.56	35.56	44.11	44.11
	MMVet	56.38	55.09	54.77	58.12	56.24	53.90	56.15	55.87	59.45	55.37
	MathVista	49.07	49.83	47.10	51.20	47.87	48.40	48.90	52.10	50.43	49.27
	MMStar	46.25	45.04	45.85	47.20	43.54	47.47	45.20	46.39	46.28	44.11
	Avg.	54.72	53.16	52.46	54.11	54.53	54.49	53.88	53.16	55.73	53.52
OCR (↑)	ChartQA	73.32	71.56	73.16	72.32	72.12	73.60	71.96	72.76	72.52	74.24
	TextVQA	66.19	64.48	64.39	65.83	66.37	64.11	63.05	65.55	65.51	64.58
	OCRBench	74.00	72.10	72.30	73.10	74.30	72.30	73.00	71.70	73.20	72.80
	Avg.	71.17	69.38	69.95	70.42	70.93	70.00	69.34	70.00	70.41	70.54
Video (↑)	MVBench	46.03	45.40	42.85	40.88	46.43	46.05	44.78	47.23	46.00	44.60
	VideoMME	52.67	50.37	47.41	46.52	51.30	52.96	49.96	51.56	51.33	51.63
	OvOBench	40.09	38.12	36.82	35.57	40.61	36.70	39.38	37.39	36.71	37.21
	Avg.	46.26	44.63	42.36	40.99	46.11	45.24	44.71	45.39	44.68	44.48
Speech (↓)	Aishell1	2.24	2.16	2.06	2.17	2.17	2.11	2.11	2.12	2.60	2.24
	Aishell2-ios	3.06	3.12	3.17	3.05	3.13	3.11	3.11	3.09	3.16	3.22
	Aishell2-android	3.13	3.17	3.22	3.09	3.10	3.12	3.23	3.17	3.23	3.30
	Librispeech-other	3.83	3.81	3.87	3.72	3.78	3.77	3.88	3.79	3.80	4.04
	Librispeech-clean	1.82	1.77	1.83	1.81	1.78	1.78	1.76	1.84	1.87	1.96
	Avg.	2.82	2.81	2.83	2.77	2.79	2.78	2.82	2.80	2.93	2.95
NLP (↑)	ARC-C	89.15	87.80	90.51	86.10	85.76	87.12	89.83	48.14	88.81	87.46
	GSM8k	81.58	82.18	85.29	80.67	80.36	83.02	79.68	83.09	81.50	81.80
	GPQA	41.92	36.36	33.84	37.88	42.93	40.40	36.36	40.40	41.92	37.37
	Avg.	70.88	68.78	69.88	68.22	69.68	70.18	68.62	57.21	70.74	68.88

guitars; in the third column sample, it identified that the brightest regions lie in the two lower sub-figures; and in the sixth column sample, it accurately recognized two bananas and four apples.

In the chart understanding task (Figure 2), the model predominantly concentrated on text-bearing regions of the table in the first two columns of samples, with negligible attention paid to blank areas. For the latter two columns, it effectively captured the extension trend of curves and key text on the coordinate axes.

Regarding the mathematics and science tasks in Figure 3, the model attended to the shapes of figures in the first two columns of math samples, prioritizing the positions of vertices and numerical values. In science tasks, it exhibited high attention to the entire circuit image (third column) owing to the sample's comparative complexity; for the subse-

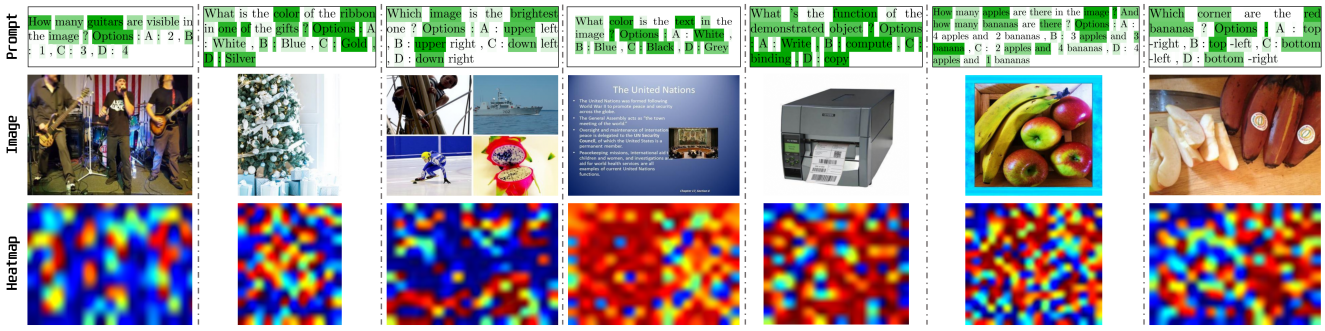

Figure 1. Token importance heatmaps for more text–image samples in general tasks. In heatmaps, redder regions indicate higher importance; bluer regions indicate lower importance. In prompts, areas with deeper green indicate high importance score.

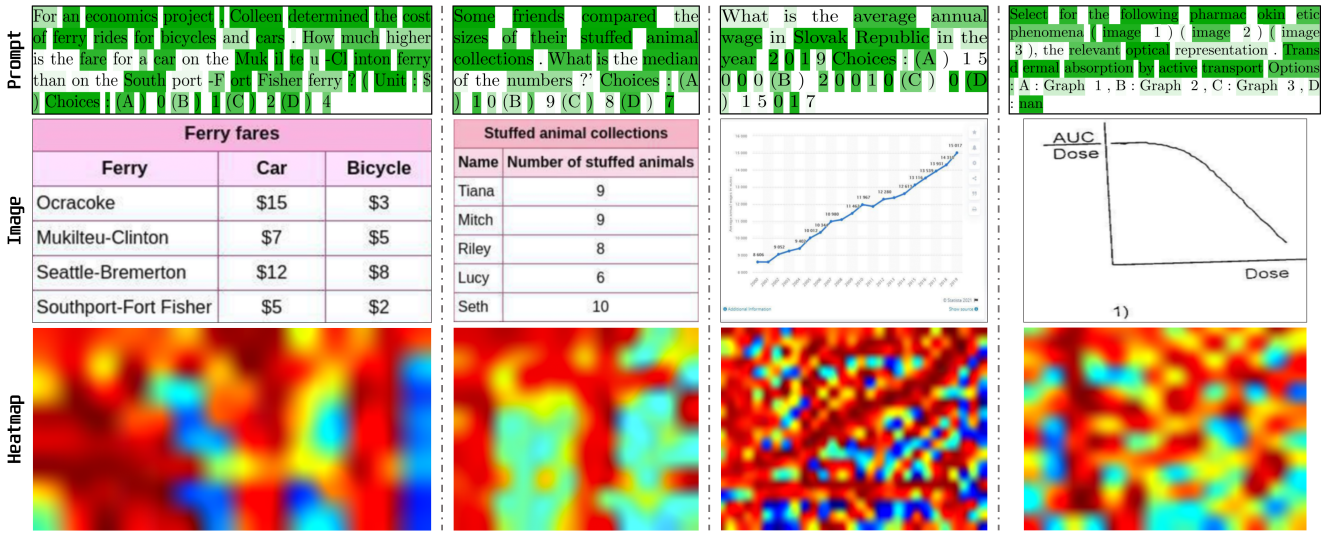

Figure 2. Token importance heatmaps for text–image samples in chart recognition task.

quent simpler leaf image, it maintained a sharp focus on the target object. In the final two geography samples, the model noted the number of targets in sub-figures and the process transformations illustrated.

Furthermore, across the three figures above, the model's attention to different tokens in the text modality is highly discriminative. It shows extremely high attention to terms denoting key targets, while non-critical terms such as "the" and "a" receive generally low attention.

Finally, we present additional video–text examples in

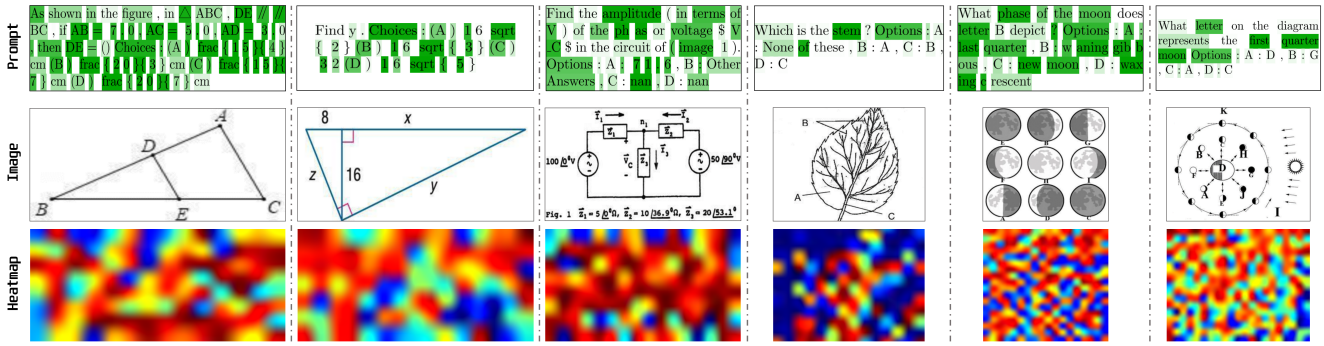

Figure 3. Token importance heatmaps for text–image samples in math and science tasks.

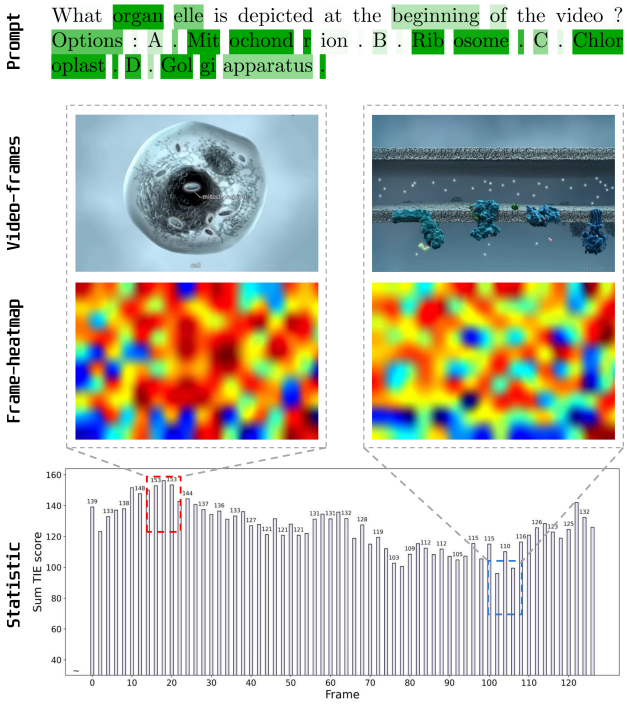

Figure 4. Token importance analysis for text–video samples.

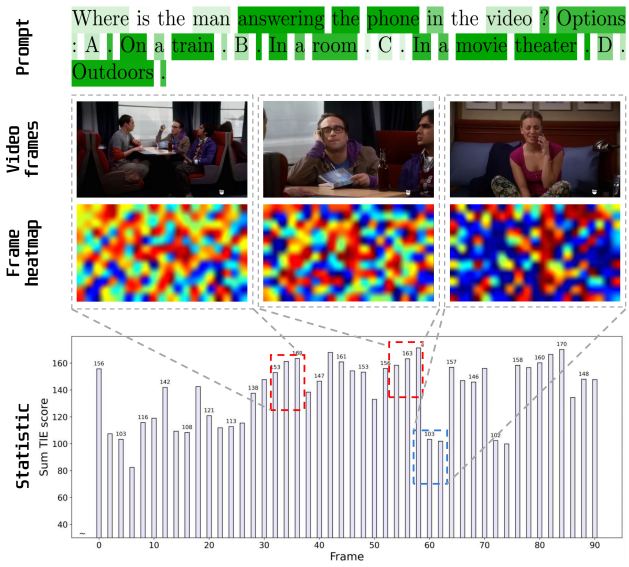

Figure 5. Token importance analysis for text–video samples.

Figure 4 and Figure 5 that further illustrate our model's temporal adaptivity. On frames highly relevant to the question, the model assigns higher importance scores and consequently activates significantly more experts; in contrast, less informative frames receive lower importance scores and exhibit reduced average expert activation. This dynamic, content-aware routing enables the model to allocate computational resources more effectively by prioritizing semantically critical information across time.

In summary, it is clear that our model can devote more computational resources to the critical content across different modalities, thus yielding promising performance.