Do All Individual Layers Help? An Empirical Study of Task-Interfering Layers in Vision-Language Models

Supplementary Material

A. Models and Benchmarks

We present all the models used in our experiments in Table 4, and list all the benchmarks we utilize in Table 5.

In this work, all datasets are evaluated using accuracy as the sole metric. The majority of datasets: MMStar, MMMU, SEEDBench, MMBench, and ScienceQA are multiple-choice question (MCQ) benchmarks, where the model's predicted option is extracted from its output and matched against the ground truth. MathVista, while formulated as a vision-question-answering (VQA) task, also employs direct string matching between generated responses and reference answers in its official evaluation, ensuring consistency in the metric across all tasks.

Specifically, MMStar is a comprehensive benchmark with 250 balanced samples across six core capabilities: Coarse Perception, Fine-grained Perception, Instance Reasoning, Logical Reasoning, Math, and Science & Technology. MMBench contains 2,974 MCQs assessing a wide range of abilities, including Coarse Perception, Fine-grained Perception (both single and cross instance), Instance Reasoning, Logic Reasoning, Attribute Reasoning, and Relation Reasoning. MMMU spans 30 disciplines, including Art & Design, Business, Science, Health & Medicine, Humanities & Social Sciences, and Engineering, covering 183 subfields with 30 types of heterogeneous images (e.g., charts, diagrams, maps, tables, musical scores, chemical structures), focusing on advanced perception and reasoning with domain-specific knowledge. SEED-Bench comprises 19,000 human-annotated MCQs, covering 12 evaluation dimensions, including image understanding. MathVista is a challenging benchmark that combines diverse mathematical and visual reasoning tasks, consisting of 6,141 examples drawn from 28 existing multimodal math-related datasets and three newly curated datasets. Finally, ScienceQA consists of 21,208 multimodal science questions collected from elementary and high school curricula. This diverse and rigorous selection of benchmarks enables a comprehensive evaluation of task-specific abilities under a unified accuracy metric.

B. Experimental Details

In our TaLo experiments, the procedure for identifying the optimal intervention layer begins with task definition and sample preparation. We first identify the target task according to the dataset's metadata, after which we draw samples from a probing pool held entirely separate from the final test set to prevent any data overlap.

Our identification process follows an iterative pipeline. We first establish a baseline performance by evaluating the unmodified model on an initial set of probing samples. If the baseline accuracy reaches 100%, the sample set is considered uninformative and is discarded; a new set is then drawn from the probing pool, and the baseline is re-evaluated. Once the baseline is established, we proceed with a systematic, layer-by-layer parameter intervention and measure the performance gain for each. If a unique layer yields the maximum positive gain, it is designated as the optimal target. In cases where multiple layers tie for the best performance or no layer produces a positive gain, we initiate a multi-round, augmented sampling strategy to resolve the ambiguity. This involves supplementing the set with an additional $shot/2$ samples for re-evaluation, followed by a further $shot/4$ samples if the tie persists. Should a unique optimal layer still not be identified after these two rounds, we select the layer with the highest index among the final candidates to ensure robustness[15, 31, 43, 52].

For our fine-tuning experiments, we employ two parameter-efficient fine-tuning (PEFT) methods: LoRA (Low-Rank Adaptation) and its variant OFT (Orthogonal Finetuning). All experiments are conducted on the LLaVA-Next-8B model. In the case of LoRA, we set the rank $r = 8$ and scaling factor $\alpha = 16$, and apply the adapter modules to all linear projections in both the language and vision pathways. This full-architecture adaptation strategy ensures comprehensive alignment of both visual and textual representations during fine-tuning. To ensure that the model truly understands the knowledge underlying the questions during fine-tuning, rather than simply memorizing the options, we format the answers as "option + option content". This approach helps the model learn the specific meaning of each option and its relationship to the question.

For model merging experiments, the LLM used for merging is DeepSeek-R1-Distill-Llama-8B[11], with a fusion coefficient λ of 0.9. However, as shown in Table 3, the effectiveness of model merging is highly sensitive to both the choice of the external LLM and the target task, suggesting that its performance is not robust across configurations and requires careful, task-specific tuning.

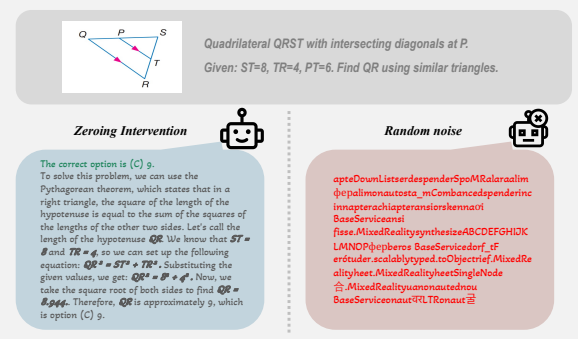
Name	Size	Huggingface ckpt
LLaVA-Next-LLaMA3 [24]	8B	`llava-hf/llama3-llava-next-8b-hf`
Qwen2-VL [47]	2B	`Qwen/Qwen2-VL-2B-Instruct`
InternVL2 [7]	26B & 40B	`OpenGVLab/InternVL2-40B`

Table 4. Details of the models used in our experiments.

Benchmark	Category	Huggingface URL
MMStar [4]	MCQ	`Lin-Chen/MMStar`
MMBench-EN [27]	MCQ	`lmms-lab/MMBench`
MMMU-VAL [54]	MCQ	`MMMU/MMMU`
ScienceQA-VAL [28]	MCQ	`derek-thomas/ScienceQA`
MathVista-MINI [29]	VQA	`AI4Math/MathVista`
SEEDBench-IMG [23]	MCQ	`lmms-lab/SEED-Bench`

Table 5. Details of the benchmarks used in our experiments.

Figure 5. Qualitative case study on random noise intervention.



C. Additional Experimental Results and Analysis

C.1. Specific clustering details

Table 11 provides the comprehensive list of tasks included in each of the seven clusters. From the table, the clustering appears to meaningfully group tasks by functional similarity. For instance, Cluster 1 brings together `numeric commonsense`, `arithmetic reasoning`, and `math word problem`—all clearly centered on numerical understanding and calculation. This suggests the method successfully identifies and isolates quantitative reasoning as a coherent capability.

Similarly, Cluster 3 stands out by grouping domain-specific scientific tasks—`Astronomy`, `Chemistry`, and `Scientific Reasoning`—into a unified theme, reflecting shared reliance on formal scientific knowledge.

C.2. Ablation on Intervention Component

In the main paper, we defines layer intervention as the modification of self-attention modules. To justify this choice, we conducted comprehensive ablation experiments applying the same intervention to the MLP modules.

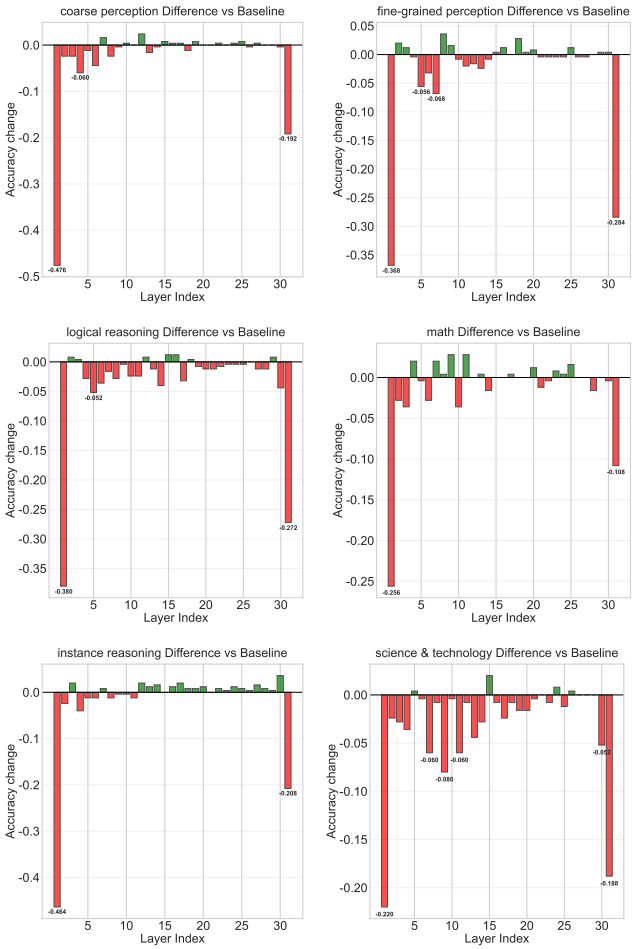
As shown in Figure 6, the intervention effect on MLP modules demonstrates a strong position-dependent sensitivity and a pattern fundamentally different from that of self-attention modules. We found that intervening on the early layers or the final layer of the model results in a catastrophic performance collapse across all tasks. As for the vast majority of middle layers, while intervening on the MLP module does not cause model collapse, showing a degree of robustness, it also fails to produce any significant performance gains. This finding justifies our decision to focus this study on self-attention modules, as they clearly exhibit the significant performance-boosting Task-Interfering Layer phenomenon we aim to investigate. Consequently, all subsequent analysis and the proposed TaLo method operate exclusively on these self-attention modules.

C.3. Additional Analysis of TaLo

Ablation study of the Intervention Method While earlier results (Section 3.2) suggest that different intervention types induce similar layer-wise trends overall, their impact can vary subtly across individual tasks. To examine these influences in a controlled yet representative setting, we turn to MMStar and MMBench: two balanced benchmarks that aggregate data from multiple sources, offering moderate task diversity and comprehensive coverage. It avoids the extremes of highly specialized or overly narrow datasets, making it well-suited for studying how these factors interact with TaLo's behavior.

In addition to parameter zeroing and uniform scaling,

Figure 6. Accuracy change after intervening on MLP layers



we also explored replacing layer weights with their mean value and random noise. However, experiments show that injecting random noise severely destabilizes the model, effectively erasing the learned representations (as shown in Figure 5). The outputs become incoherent, often degenerate, with little connection to the input. This suggests that the pre-trained parameters, even when suboptimal for a specific task, still encode essential structural and semantic priors critical to model functionality.

Given this breakdown in basic competence, we focus our analysis on milder, more controlled interventions: zeroing, uniform scaling, and mean replacement, which preserve the model's foundational knowledge while allowing targeted modulation. These methods strike a better balance between perturbation and stability, enabling meaningful study of layer-wise task adaptation without collapsing overall performance.

To further examine the impact of three intervention types, we conduct ablation studies across a wide range of tasks on two benchmark datasets, using a consistent 10-shot setting. As shown in Table 6,7, we observe that zeroing and uniform scaling yield comparable effects, with zeroing achieving better average performance across tasks. In contrast, mean replacement consistently leads to inferior results.

This observation aligns with our earlier findings in Section 3.2, where both scaling and zeroing exhibited similar layer sensitivity patterns. While subtle differences may arise in specific contexts, which depend on model architecture or task nature.

In practice, the choice between scaling and zeroing can depend on task-specific behavior or implementation simplicity. Both support effective plug-and-play adaptation. TaLo's strength appears to lie not in the intervention itself, but in the strategic selection of where and when to apply it. It is the layer not the operation seems to be the more decisive factor.

Multi-Layer Interventions To complement the main results based on single-layer intervention, we conduct a systematic study of two-layer TaLo interventions. For each task, we first identify the optimal single layer using the standard TaLo procedure. We then fix this layer and iteratively apply a second zeroing intervention to every other layer in the LLM backbone, measuring the resulting performance change while keeping all other components unchanged. This yields a full pairwise intervention matrix for each task, from which we select the top-performing two-layer combination.

Due to the quadratic growth in computational cost with model depth, we limit our exploration to two-layer combinations as a tractable proxy for higher-order interactions. The results consistently show limited or no gain from adding a second intervention, reinforcing the sparsity of task-interfering layers observed in the main paper.

Robustness on layer selection To demonstrate that the robustness of our layer selection is not limited to a specific domain, we extended the bootstrapping analysis (N=50) to two distinct task categories across three few-shot settings. As illustrated in Figure 7, we observe that the layer selection distribution remains highly concentrated across all tasks, confirming that TaLo consistently localizes a stable region of interference regardless of the task type or sampling variations. Crucially, the distinct locations of these interference regions across different tasks powerfully validate that TaLo captures meaningful, task-specific functional conflicts inherent to the model's internal representations, rather than merely identifying universally redundant layers.

More Experimental Results of TaLo As evidenced by the comprehensive results in Table 9 10, which encompasses evaluations on MMBench and ScienceQA, TaLo

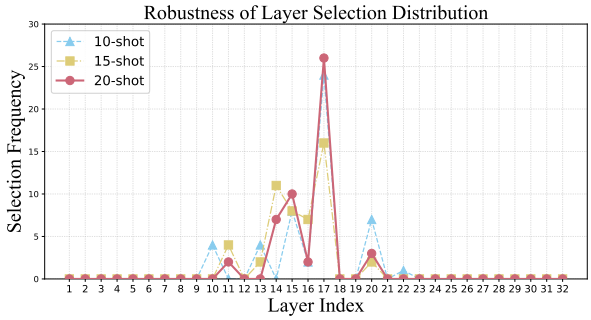
Table 6. Results of TaLo on *LLaVA-Next-Llama3-8B* under different intervention methods. The last column (**Avg**) reports the mean change across tasks. ✗ indicates the method failed to find the Task-Interfering layer.

Intervention	MMStar						Avg
	Coarse perception	Fine-grained perception	Instance reasoning	Science & technology	Logical reasoning	Math	
Zeroing	63.8 ↓2.9	41.9 ↑0.9	57.6 ↑3.8	32.9 ↑1.0	38.6 ↓3.8	31.0 ↑2.9	0.32↑
Uniform scaling	66.2 ↑2.4	41.0 ↓0.4	51.0 ↓2.8	24.8 ↓5.2	40.0 ↓1.0	34.3 ↑3.8	0.53↓
Mean replacement	✗	34.8 ↓6.2	52.9 ↓0.9	27.6 ↓3.4	41.9 ↑1.4	30.5 ↑1.0	1.14↓

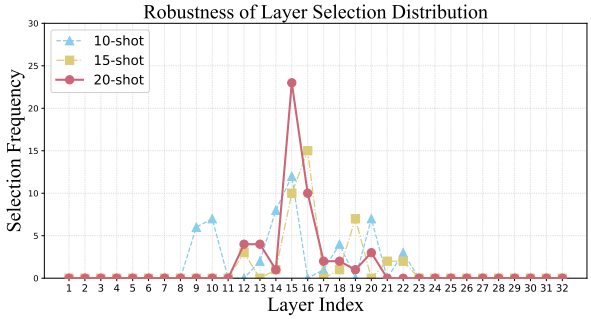
Table 7. Results of TaLo on *LLaVA-Next-LLaMA3-8B* under different intervention methods (*Structuralized i-t understanding* stands for Structuralized image text understanding).

Intervention	MMBench					Avg
	Physical property reasoning	Structuralized i-t understanding	Attribute recognition	Celebrity recognition	Image emotion	
Zeroing	55.3 ↑7.8	55.8 ↑0.8	65.2 ↓3.1	68.5 ↓1.4	65.6 ↑0.6	0.94↑
Uniform Scaling	57.0 ↑5.0	55.8 ↑0.8	71.0 ↑0.5	67.7 ↓2.8	65.6 ↓0.6	0.58↑

Figure 7. Layer Selection's Robustness Analysis.



(a) Analysis on Math task



(b) Analysis on Instance-reasoning task

demonstrates a robust ability to enhance performance across a wide spectrum of tasks. Nonetheless, the magnitude of improvement is observed to be more constrained and in some cases even decreases for particularly challenging categories, such as those involving complex multi-step reasoning, detailed visual attribute discrimination. This pattern suggests that while our proposed layer-level intervention provides an effective mechanism for task adaptation, its efficacy is bounded by the underlying capabilities of the pre-trained model. Performance plateaus or regressions in these demanding scenarios likely point to limitations that are architectural or data-based in nature, which might be addressed in future work by integrating stronger inductive biases or auxiliary knowledge sources.

C.4. Empirical Validations of Task-Interfering Layers

The accuracy change heatmaps across multiple models (*LLaVA-Next-Llama3-8B*, *Qwen2-VL-2B*, and *InternVL2-40B*) under different intervention strategies are shown in Figures 9 to 20. Evaluated on diverse benchmarks, these heatmaps reveal consistent patterns of layer-specific performance gains, forming the core empirical basis for the task-interfering layers phenomenon. Rather than isolated anomalies, the results suggest a systemic trade-off in how individual layers support competing task demands, observable across model scales and architectures.

We observe that Task-Layer Interaction Vectors exhibit significant morphological differences across VLMs, even for identical tasks. We argue this is not a limitation but a critical, expected finding. Since foundation models diverge in architectures, pre-training data, and objectives, they develop unique paths of internal functional specialization. Consequently, task interference is inherently a model-specific phenomenon; a layer hindering performance in one model may not do so in another. This variability renders universal, static removal strategies impractical. It strongly underscores the necessity of our adaptive approach, TaLo. By utilizing a small probing set, TaLo effectively handles this heterogeneity, dynamically identifying and bypassing the specific interfering layers unique to each model's representational space.

Table 8. Results of TaLo on *LLaVA* under two-layer intervention (10-shot). '✗' marks cases where a second Task-Interfering Layer could not be identified. Details of the MMStar are provided in Appendix A.

Metric	MMStar					
	CP	FP	IR	S&T	LR	Math
Task-Interfering Layer	L1, L6	L15, L29	L11, ✗	L31, ✗	L1, L8	L6, L13
Performance (two layers)	61.9 ↓3.8	40.0 ↓0.5	56.2 ↑1.0	31.0 ↑0.5	37.1 ↓3.8	25.7 ↓3.8
Performance (single layer)	63.8 ↓2.9	41.9 ↑0.9	57.6 ↑3.8	38.6 ↓3.8	31.0 ↓2.9	32.9 ↑2.5

Figure 8. Qualitative Case Studies Illustrating the Effects of Layer Zeroing on LLaVA-Next's Reasoning. The figure* presents three comparative examples of the model's reasoning process before (base model) and after the intervention of a specific task-interfering layer.

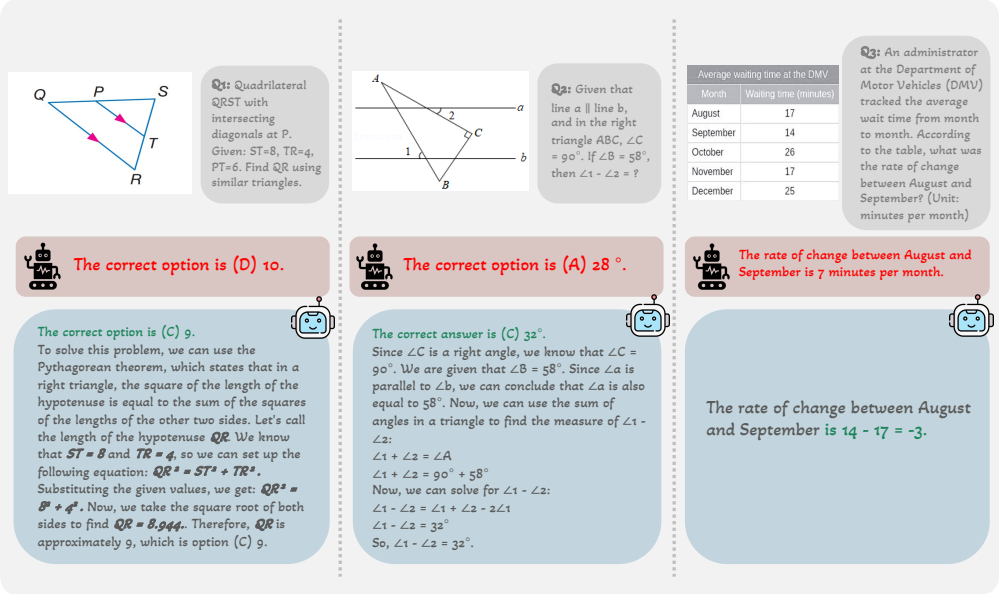


Table 9. Additional results of TaLo on *LLaVA-Next-LLaMA3-8B*. Here, *future prediction* is annotated as F-P, and *function reasoning* is annotated as F-R.

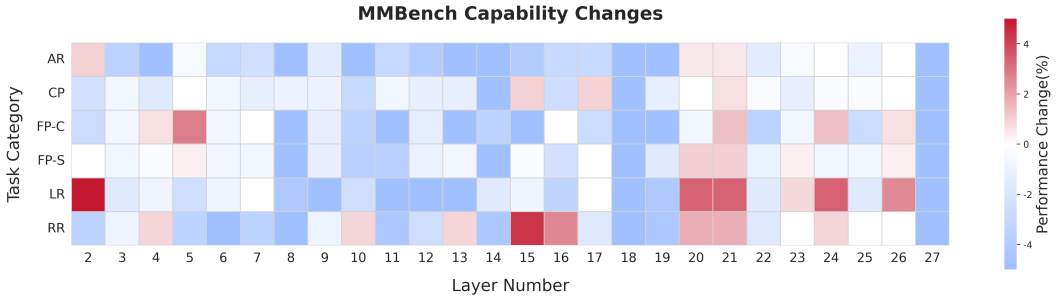
Model	Shots	MMBench			ScienceQA			
		F-P	OCR	F-R	Ecological interactions	The Americas: Geography	Oceania: Geography	Geography
LLaVA	**10 shots**	43.3 ↓3.3	69.8 ↓2.6	68.6 –0.0	17.6 –0.0	30.0 ↑20.0	21.9 –0.0	41.7 ↑8.4
	15 shots	48.3 ↓1.1	71.2 –0.0	68.0 ↓1.3	29.4 ↑11.8	25.0 ↑5.0	25.0 ↑3.1	39.6 ↑4.2
	20 shots	58.6 ↑6.9	78.8 ↑3.8	71.0↑2.2	29.4↑11.8	35.0 ↑20.0	15.6–0.0	33.3 ↑2.1

Table 10. Additional results of TaLo on *Qwen2-VL-2B*, where *structuralized imagetext understanding* is annotated as S-I-U, *attribute recognition* is annotated as AR, *physical relation* is annotated as PR, and *celebrity recognition* is annotated as CR.

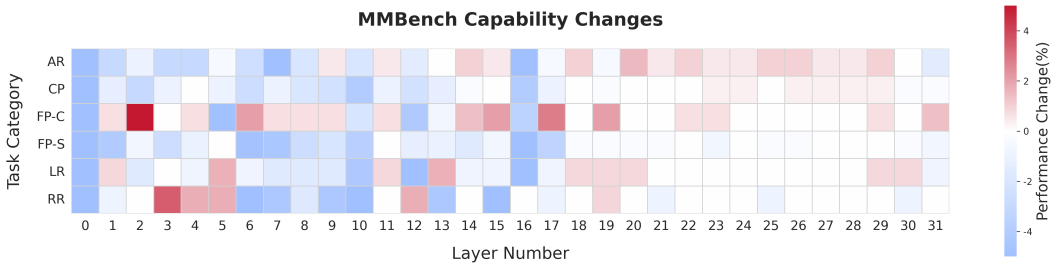
Model	Shots	MMBench				ScienceQA				
		S-I-U	AR	PR	CR	Astronomy	The Americas: Geography	Genes to traits	Solutions	Force and motion
Qwen-VL	**10 shots**	52.5 ↑3.3	72.3 ↑0.9	47.6 ↑1.6	75.0 ↑2.0	35.5 ↑3.2	25.0 ↑15.0	21.9 ↑12.5	24.4 ↑2.2	35.3 ↑5.9
	15 shots	52.3 ↑3.2	72.1 ↑2.0	50.8 ↑3.2	75.3 ↑1.5	38.7 ↑6.4	10.0 –0.0	25.0 ↑3.1	28.9 ↑8.9	29.4 ↓5.9
	20shots	48.5 ↑2.5	73.4 ↑1.1	55.6 ↑3.2	76.6 ↑1.0	32.3 ↑6.5	15.0 ↑5.0	28.1 ↑6.2	35.6 ↑15.6	52.9 ↑17.6

Cluster Category and Tasks Included	
Cluster 1 (Quantitative Reasoning)	numeric commonsense, arithmetic reasoning, geometry reasoning, algebraic reasoning, geometry problem solving, math word problem, figure* question answering, statistical reasoning, Cities, Informational texts: level 1, Particle motion and energy
Cluster 2 (Analytical Reasoning)	image_emotion, biology, engineering, public health, instance reasoning, math, geography, visual reasoning, architecture & engineering, diagnostics & laboratory medicine, electronics, psychology, maps, magnets, plant reproduction, domain-specific vocabulary, genes to traits
Cluster 3 (Scientific Knowledge)	scientific reasoning, textbook question answering, chemistry, medicine, economics, physics, sociology, art & design, science & technology, astronomy, plants, weather, fossils, thermal energy, natural resources, pharmacy, humanities & social science, literature
Cluster 4 (Integrative Reasoning)	logical reasoning, accounting, history, pharmacy, engineering practices, ecology, world religions, persuasive strategies, adaptations and natural selection, economics, sociology, humanities, design, literature, natural science theory, political history
Cluster 5 (Perceptual Categorization)	action recognition, attribute recognition, image quality, coarse perception, fine-grained perception, object localization, classification, ecosystems, force and motion, solutions, states of matter, scene understanding, OCR, image_scene, visual elements, structuralized image-text understanding, attribute comparison, celebrity recognition
Cluster 6 (Predictive Reasoning)	future prediction, identity reasoning, spatial relationship, electronics, psychology, math, music, plant reproduction, velocity & acceleration, instance interaction, physical geography, classification and scientific names, context clues, text understanding
Cluster 7 (Relational Understanding)	nature_relation, physical_relation, social relation, art theory, civil war and reconstruction, age of exploration, ancient Mesopotamia, plate tectonics, geology, animals, agriculture, ecosystems, cultural history, world religions, state capitals

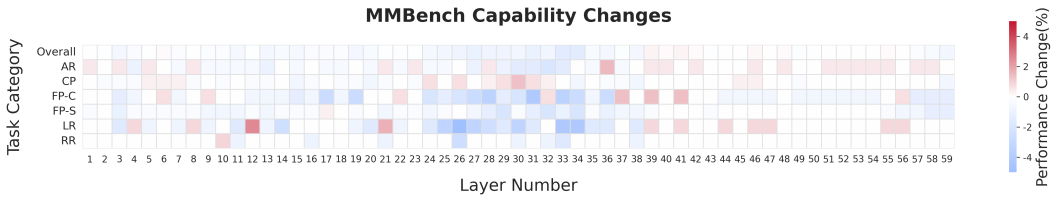
Table 11. Clusters and their included tasks from various benchmarks.



(a) Accuracy change heatmap on Qwen2-VL.



(b) Accuracy change heatmap on LLaVA-Next.



(c) Accuracy change heatmap on InternVL2.

Figure 9. Accuracy change heatmaps on MMBench (Uniform Scaling).

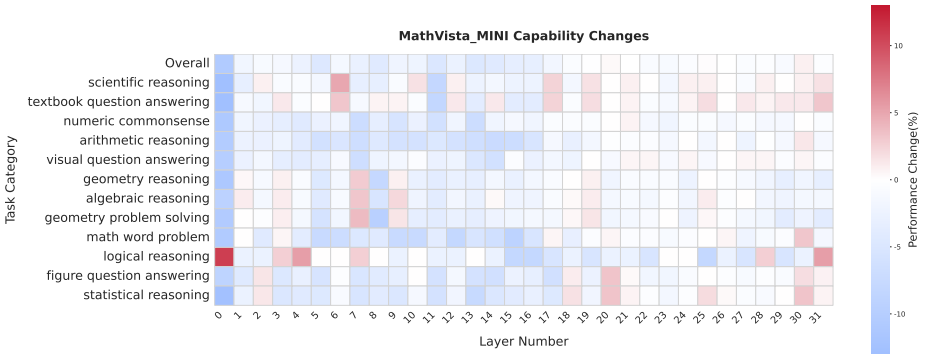


Figure 10. Accuracy change heatmap for LLaVA-Next on MathVista-MINI (Uniform Scaling).

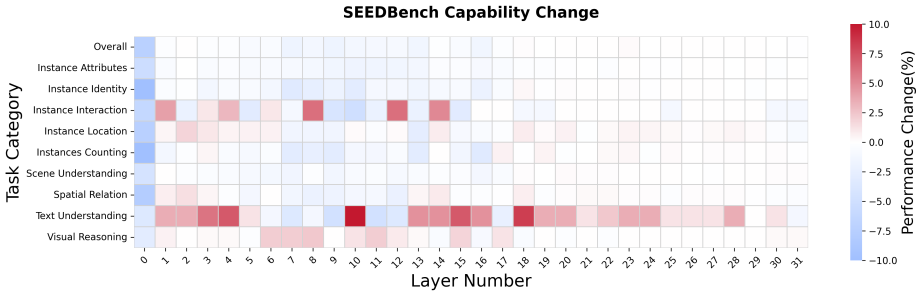


Figure 11. Accuracy change heatmap for LLaVA-Next on SEEDBench (Uniform Scaling).

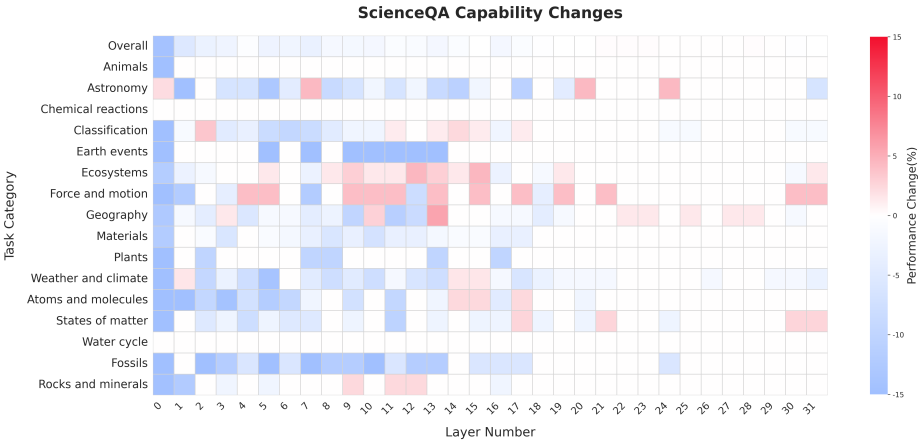
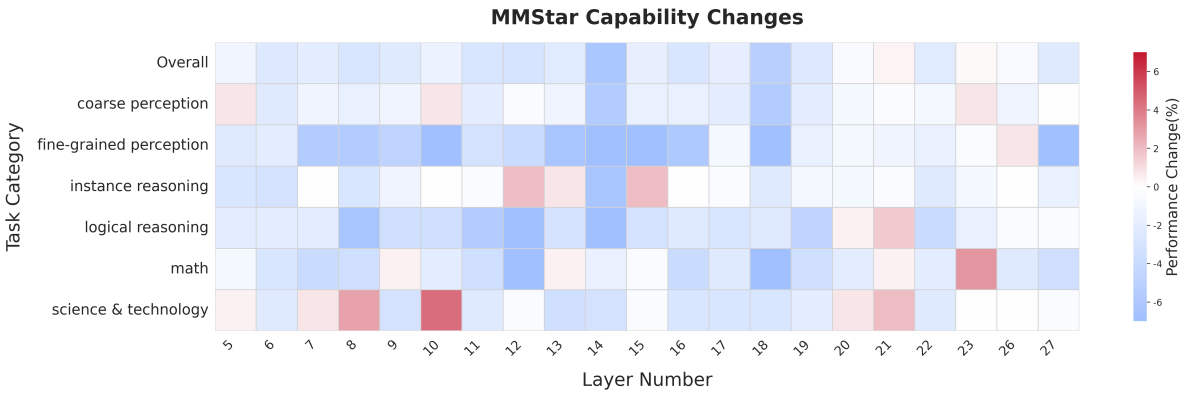
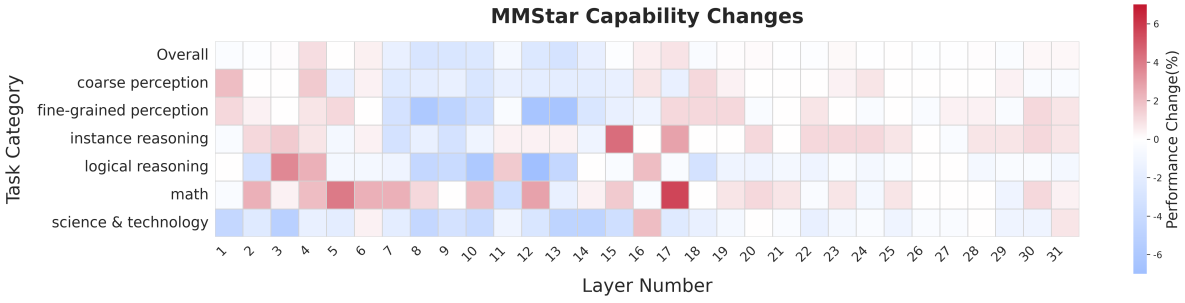


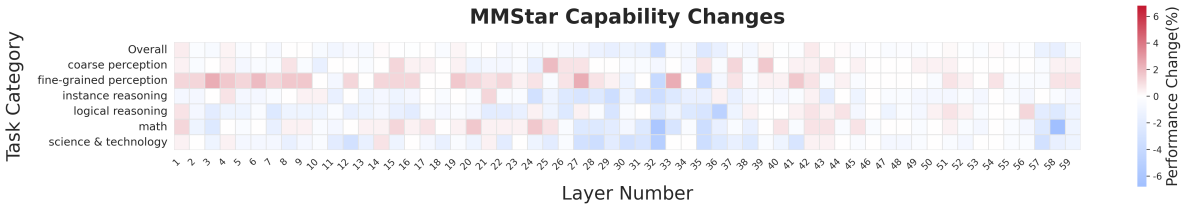
Figure 12. Accuracy change heatmap for LLaVA-Next on ScienceQA (Uniform Scaling).



(a) Accuracy change heatmap on Qwen2-VL.

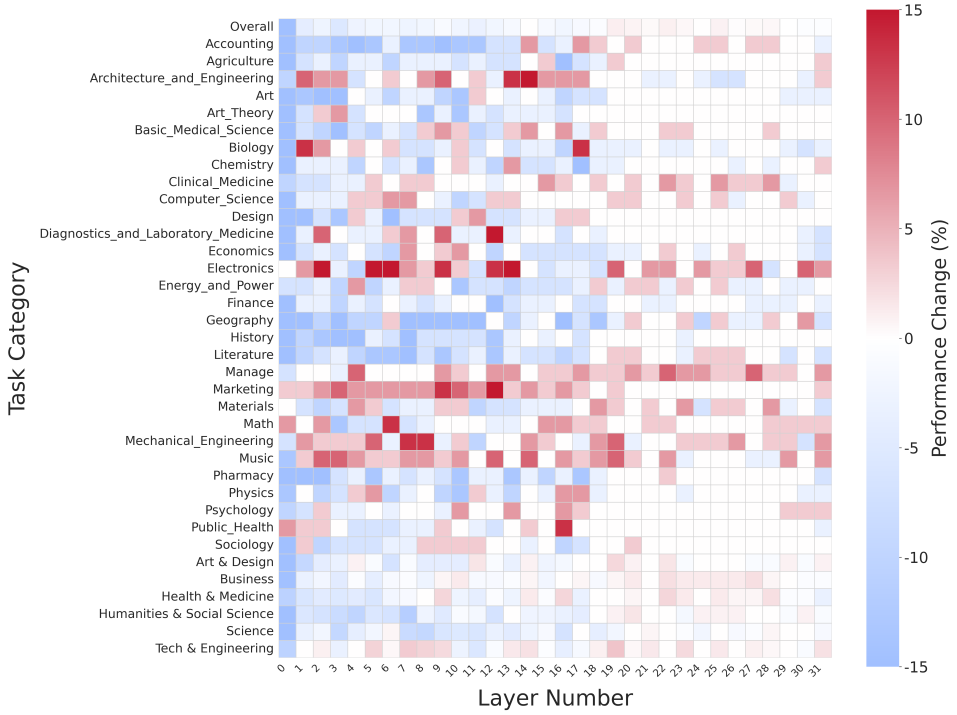


(b) Accuracy change heatmap on LLaVA-Next.

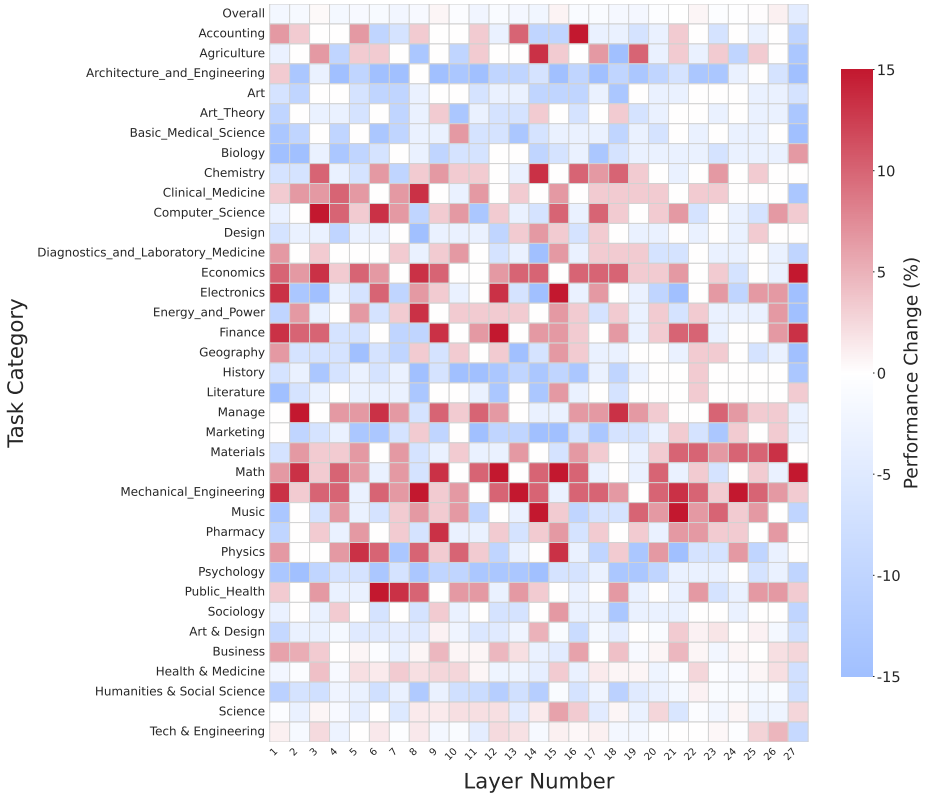


(c) Accuracy change heatmap on InternVL2.

Figure 13. Accuracy change heatmaps on MMStar (Uniform Scaling).

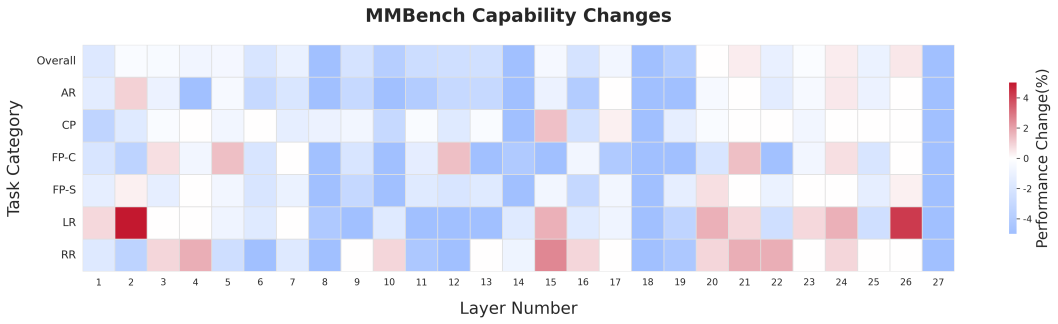


(a) Accuracy change heatmap on LLaVA-Next.

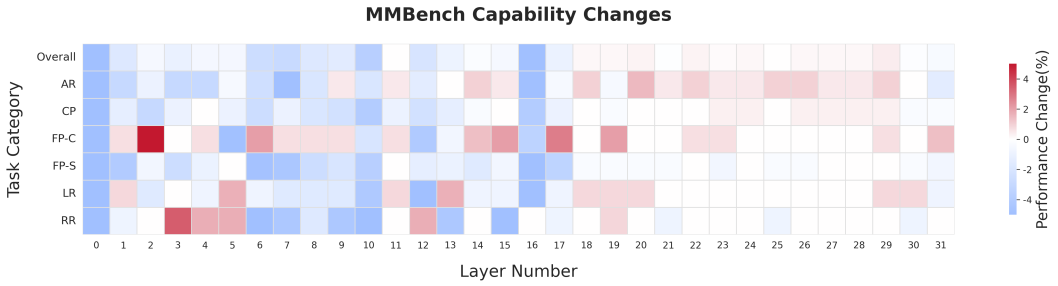


(b) Accuracy change heatmap on Qwen2-VL.

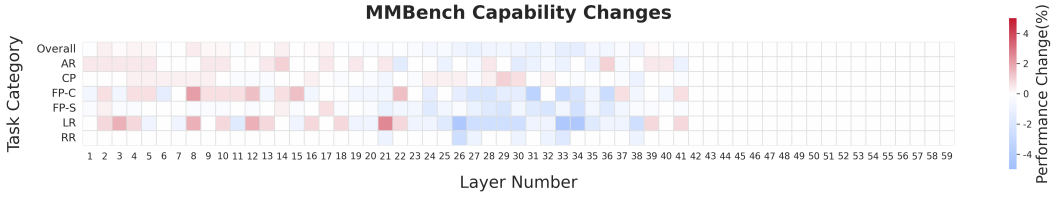
Figure 14. Accuracy change heatmaps on MMMU (Uniform Scaling).



(a) Accuracy change heatmap on Qwen2-VL.



(b) Accuracy change heatmap on LLaVA-Next.



(c) Accuracy change heatmap on InternVL2.

Figure 15. Accuracy change heatmaps on MMBench (Zeroing).

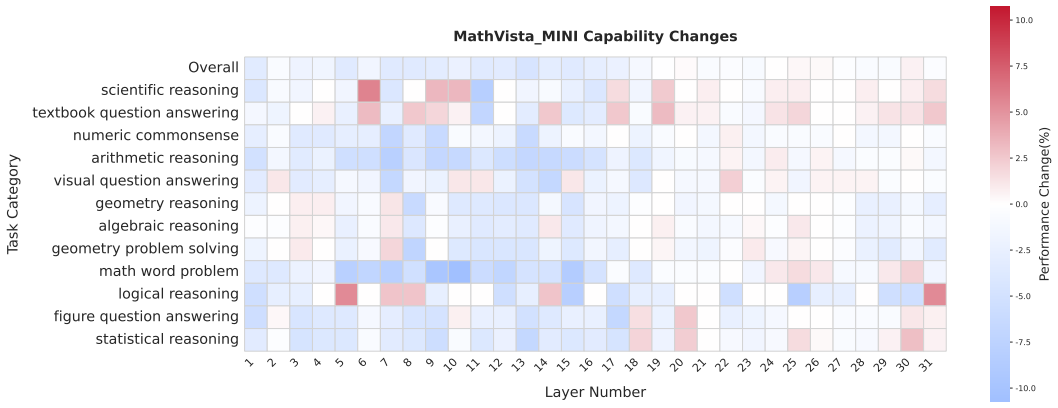


Figure 16. Accuracy change heatmap for LLaVA-Next on MathVista-MINI (Zeroing).

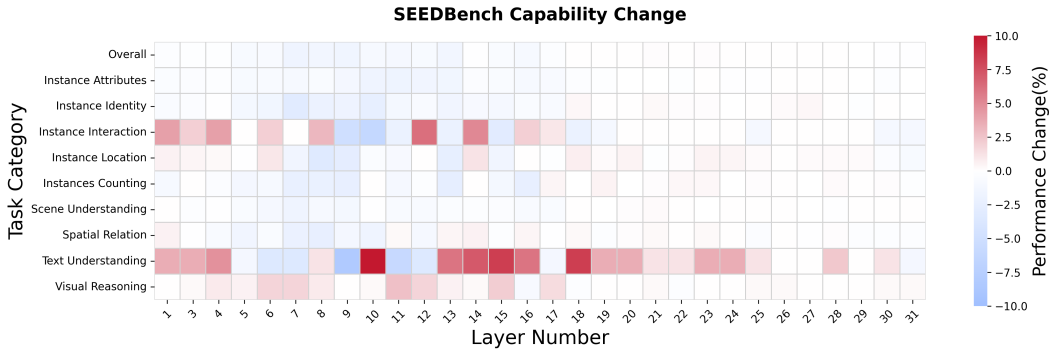


Figure 17. Accuracy change heatmap for LLaVA-Next on SEEDBench (Zeroing).

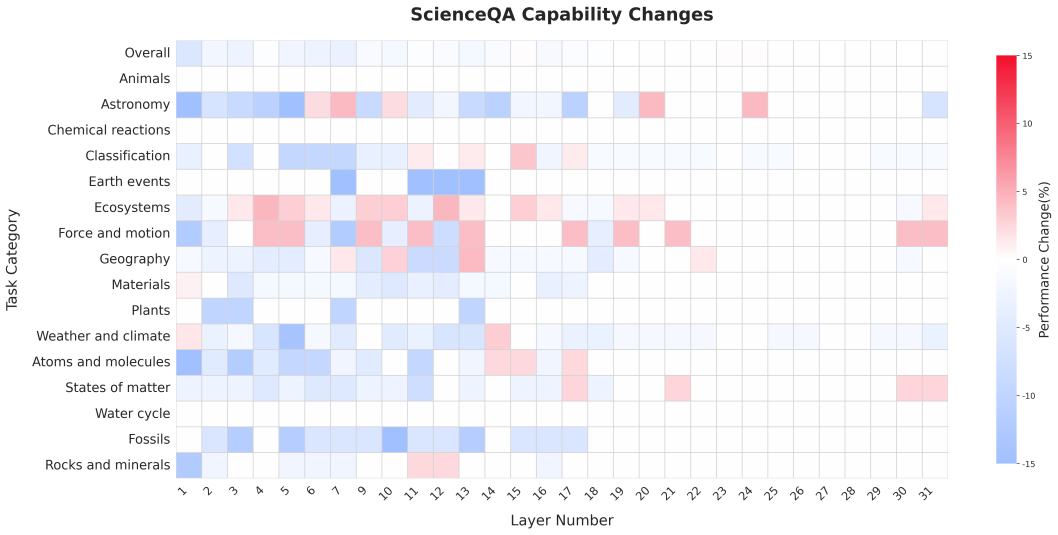
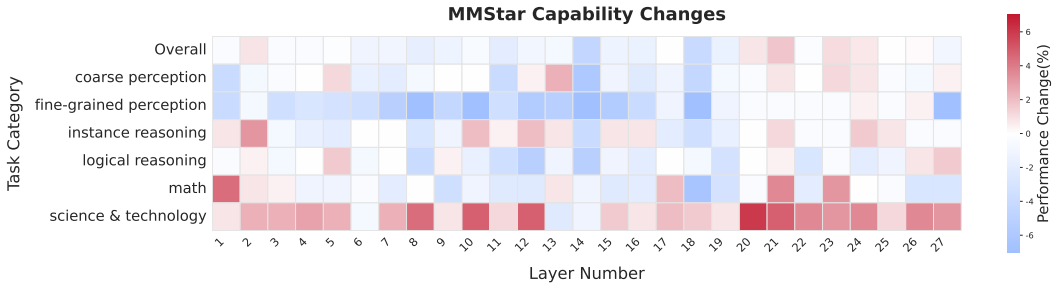
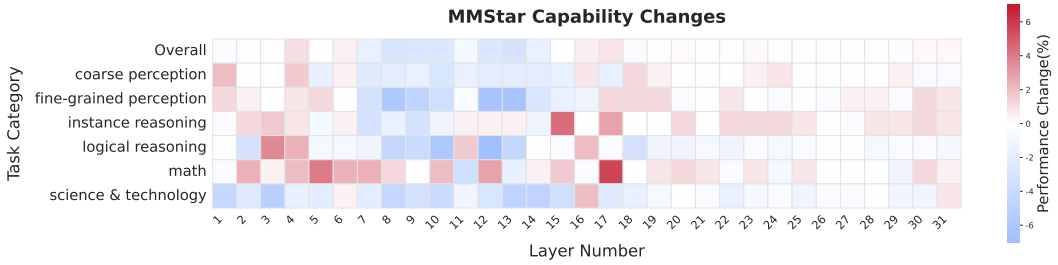


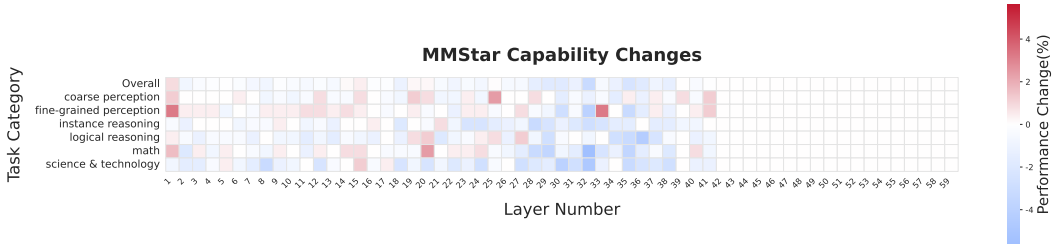
Figure 18. Accuracy change heatmap for LLaVA-Next on ScienceQA (Zeroing).



(a) Accuracy change heatmap on Qwen2-VL.

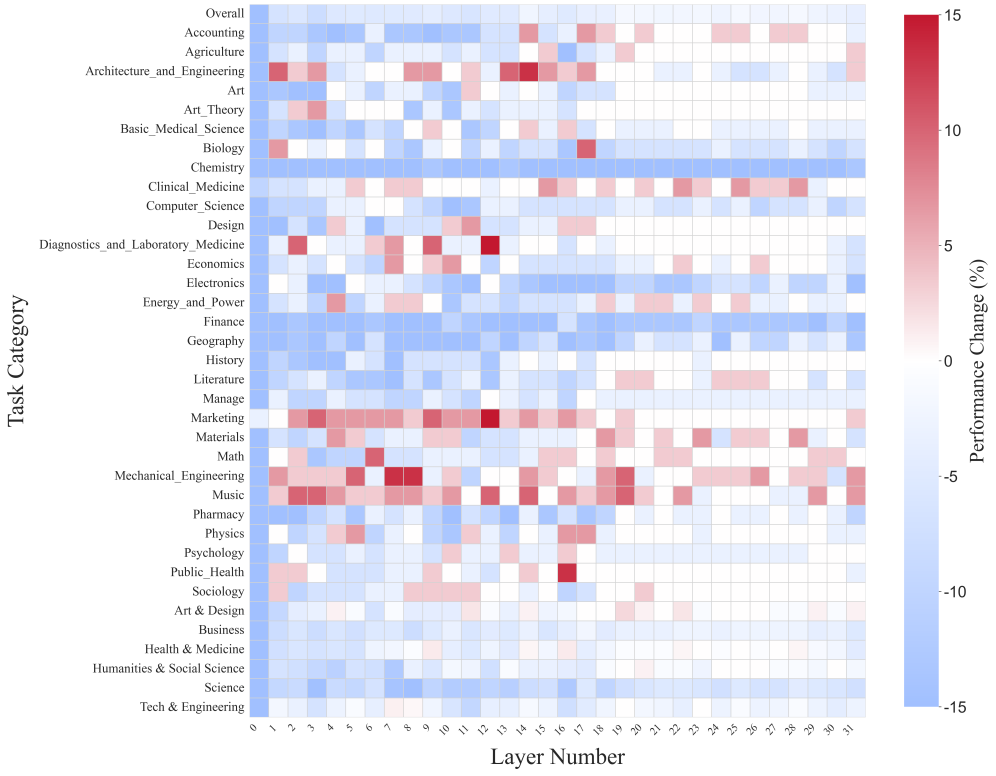


(b) Accuracy change heatmap on LLaVA-Next.

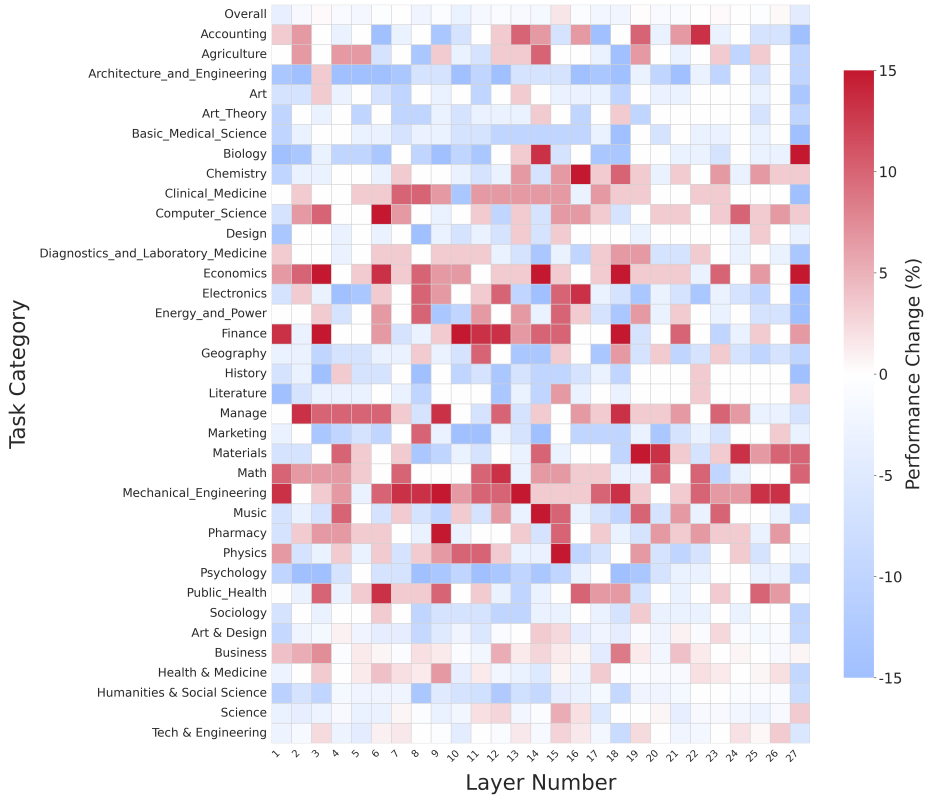


(c) Accuracy change heatmap on InternVL2.

Figure 19. Accuracy change heatmaps on MMStar (Zeroing).



(a) Accuracy change heatmap on LLaVA-Next.



(b) Accuracy change heatmap on Qwen2-VL.

Figure 20. Accuracy change heatmaps on MMMU (Zeroing).