Understanding Task Transfer in Vision-Language Models

Supplementary Material

Table of Contents

A.1. PGF Calculation and Heatmaps

We provide a pseudo-code to compute the Perfection Gap Factor in Algorithm 1. The goal of the metric is to quantify how much of the remaining achievable performance a model recovers through finetuning.

Algorithm 1 Pseduo-code to compute Perfection Gap Factor.

Require: Baseline accuracy A_{base}, finetuned accuracy A_{ft}, ceiling U, small constant ϵ
Ensure: PGF value μ

$\Delta \leftarrow A_{\text{ft}} - A_{\text{base}}$ ▷ accuracy change
gap $\leftarrow U - A_{\text{base}} + \epsilon$ ▷ remaining room to improve
$\mu \leftarrow \Delta / \text{gap}$ ▷ PGF definition
return μ

We also provide PGF heatmap for the 13 tasks with mean PGF and standard deviation, alongside transferability and malleability in Figure A.2, Figure A.3, and Figure A.4, for 3B, 7B and 32B, respectively. We note that the standard deviation remains consistently small, suggesting that the results are stable rather than driven by noise.

A.2. Accuracy Heatmaps

We also include accuracy heatmaps for all 13 tasks, reporting both the mean and standard deviation, in Figure A.5, Figure A.6, and Figure A.7 for the 3B, 7B, and 32B models, respectively. These summaries highlight how performance varies across tasks and model scales, and the accompanying standard deviations indicate the degree of variability in the underlying measurements.

A.3. Task Category Trends

In Figure A.9 and Figure A.8, we plot the negative transferability and negative malleability respectively. Unlike the positive trends, we observe a sharp negative transferability and malleability in Qwen2.5-VL-7B model. On an average across models, low-level and image-level tasks exhibit the highest magnitude of negative transferability. High-level and crop-level tasks exhibit the highest magnitude of negative malleability. Additionally, we provide the heatmaps for transferability and malleability across all the task categories in Figure A.10, Figure A.11 and Figure A.12, for model sizes 3B, 7B and 32B respectively.

A.4. Cliques across Model Sizes

Table A.1 lists the positive and negative cliques identified across all three model sizes. In addition, Figure A.13, Figure A.14, and Figure A.15 visualize the largest positive and negative clique for the 3B, 7B, and 32B models, respectively. We note that 32B variant has the largest positive clique of size 9.

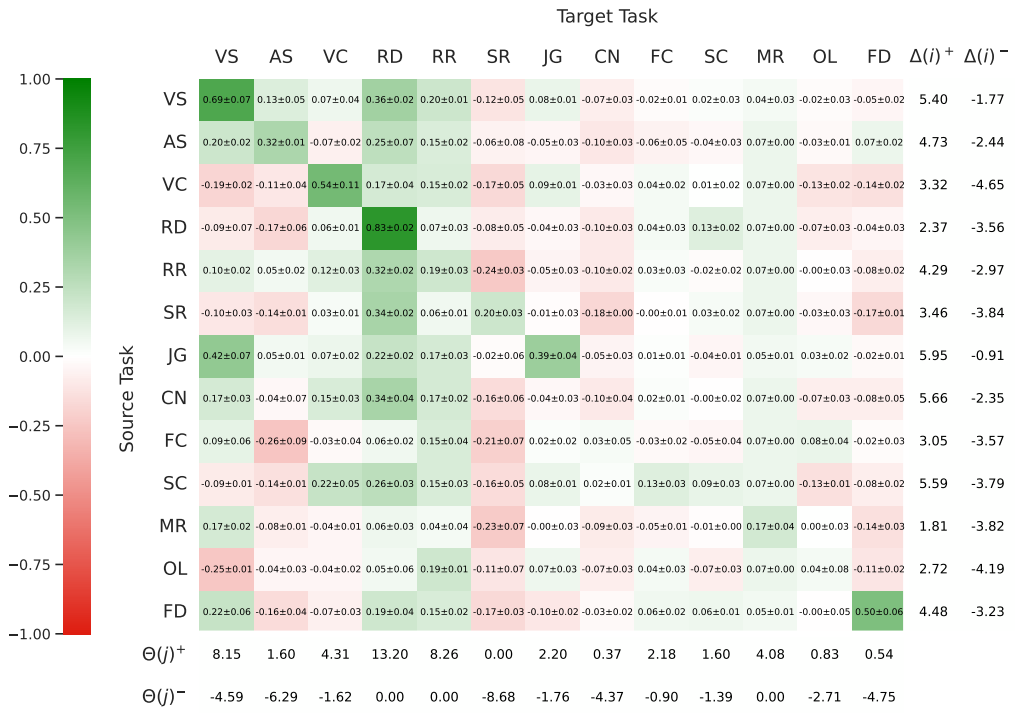


Figure A.2. PGF Heatmap for Qwen-2.5-VL 3B.

Figure A.3. PGF Heatmap for Qwen-2.5-VL 7B.

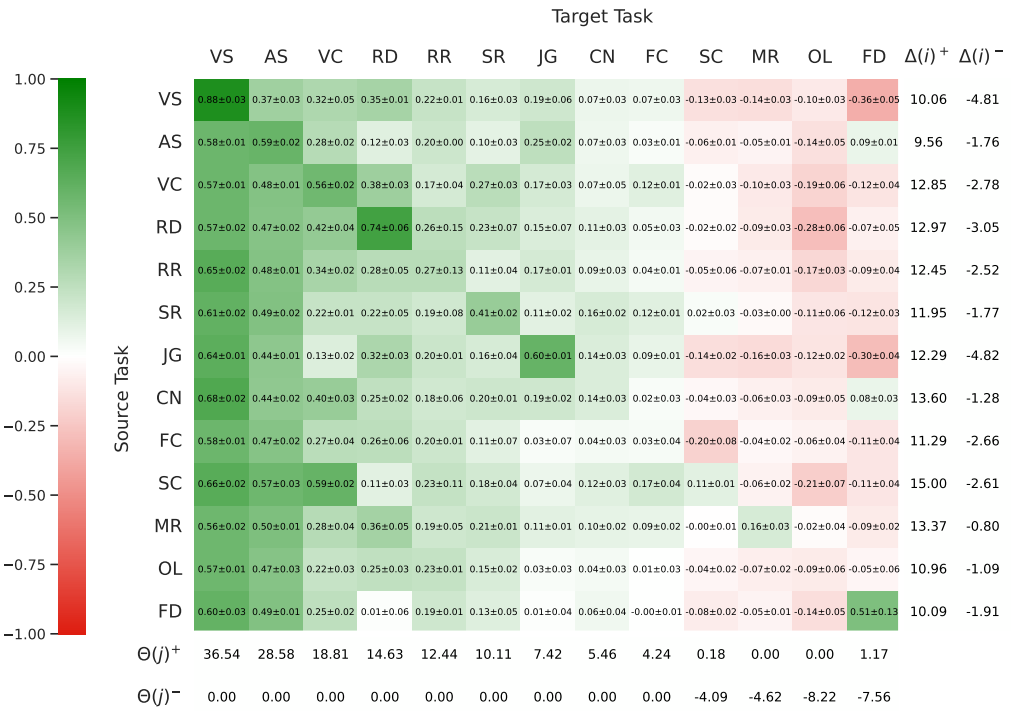


Figure A.4. PGF Heatmap for Qwen-2.5-VL 32B.

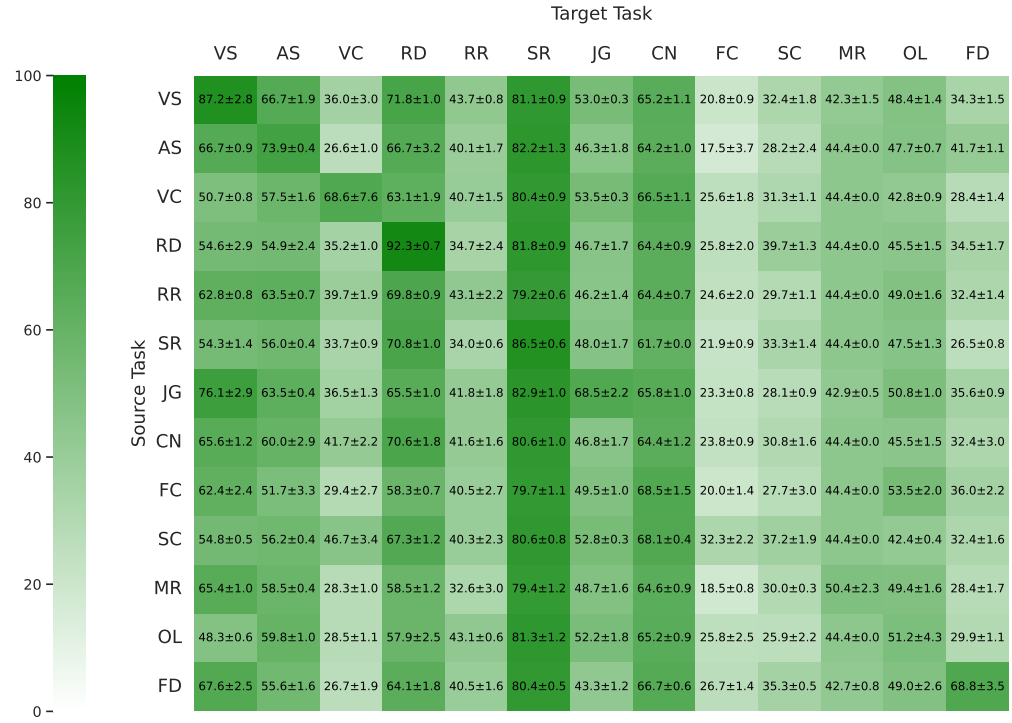


Figure A.5. Accuracy Heatmap for Qwen-2.5-VL 3B.

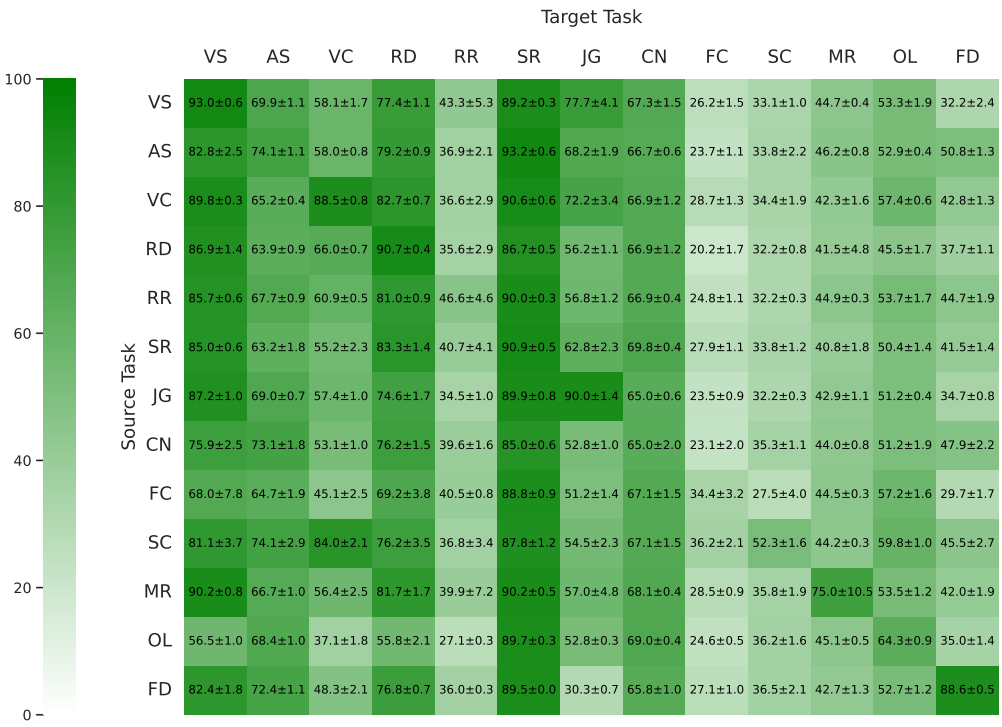


Figure A.6. Accuracy Heatmap for Qwen-2.5-VL 7B.

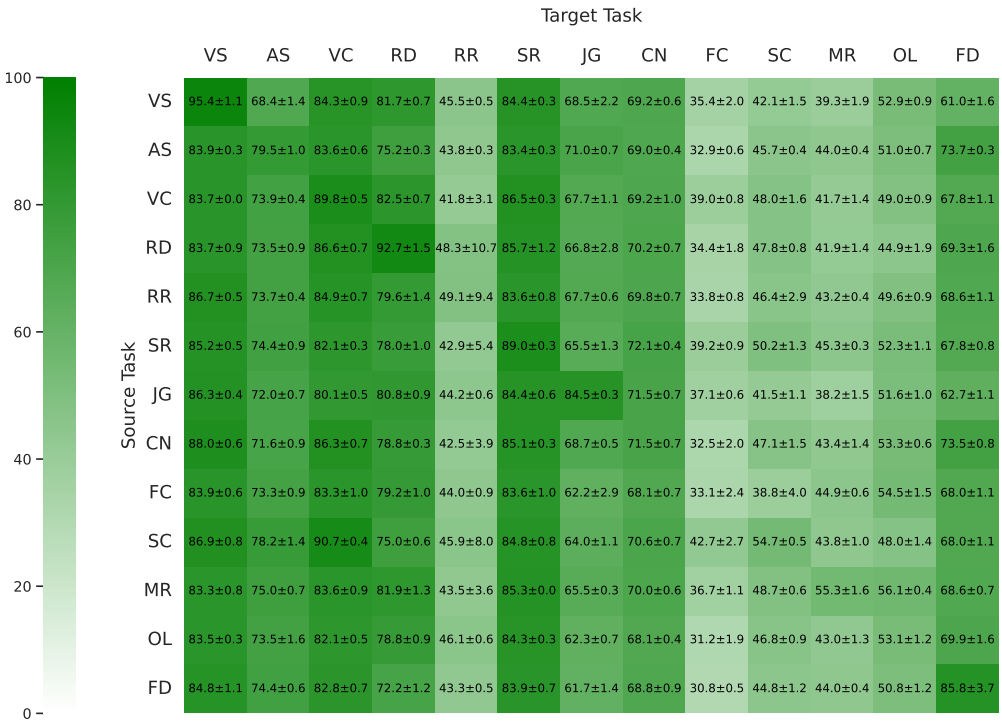


Figure A.7. Accuracy Heatmap for Qwen-2.5-VL 32B.

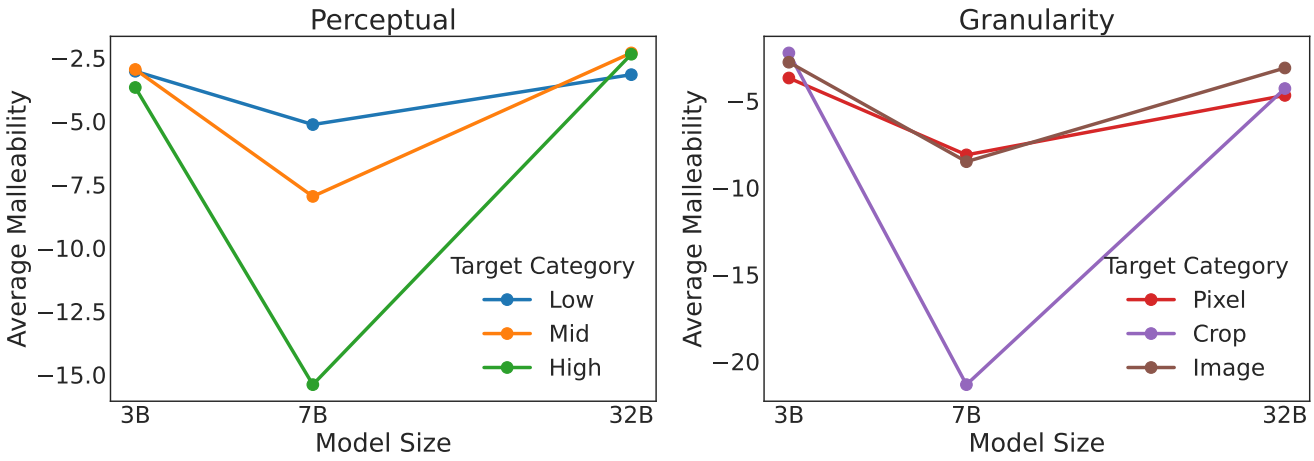


Figure A.8. Average negative malleability trends across granular and perceptual levels.

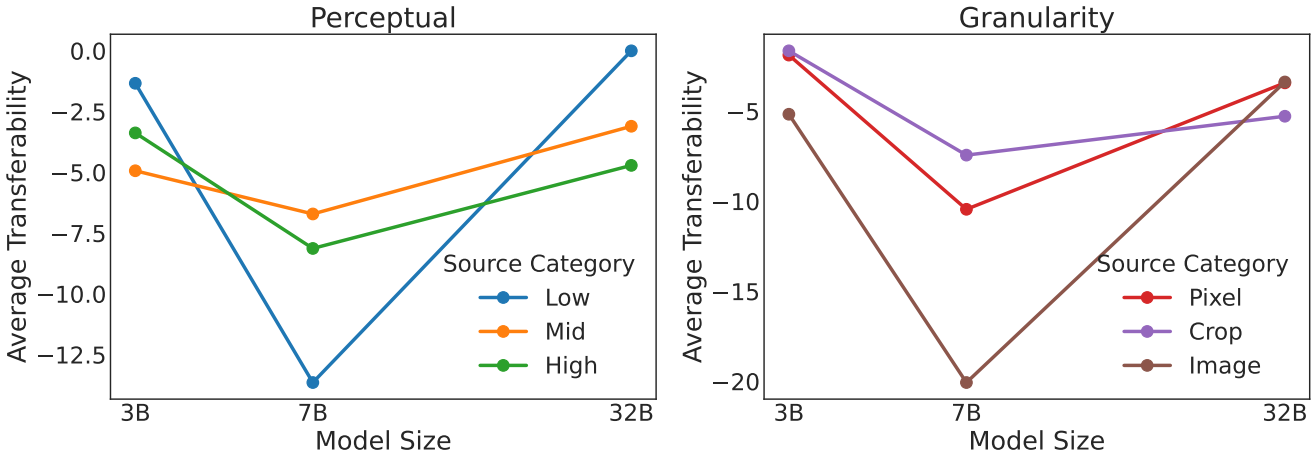


Figure A.9. Average negative transferability trends across granular and perceptual levels.

A.5. Implementation Details

All finetuning experiments are performed on 8xA100 GPUs 40GB using the opensource Qwen repository*. DeepSpeed [13] ZeRO-2 is used for Qwen-2.5-VL 3B and 7B, while DeepSpeed [13] ZeRO-3 is used for Qwen-2.5-VL 32B, all with mixed-precision. Batch size is set to 16, weight decay as 0 and warmup ratio of 0.03 with cosine decay learning rate scheduler. For finetuning, LoRa rank is set to 8 and α is set to 16 for all tasks. Task-wise training details are mentioned in Table A.2. We utilize the GPT-4.1 model for extracting responses from model responses and the evaluation is performed using the official code provided by the BLINK benchmark†.

A.6. Effect of Training Steps on PGF

In Figures A.16, Figure A.17 and Figure A.18, we examine the impact of finetuning steps on transferability using Qwen2.5-VL-3B. These heatmaps show that with increasing number of steps, average transferability increases monotonically. This behavior is expected as additional optimization amplifies the model's deviation from the original checkpoint, strengthening transfer signals. While the absolute PGF values change with training duration, the qualitative structure of the transfer patterns remains stable across ablations, indicating that the relationships among tasks are largely preserved even under longer finetuning.

*https://github.com/QwenLM/Qwen3-VL
†https://github.com/zeyofu/BLINK_Benchmark
‡https://huggingface.co/datasets/keremberke/painting-style-classification

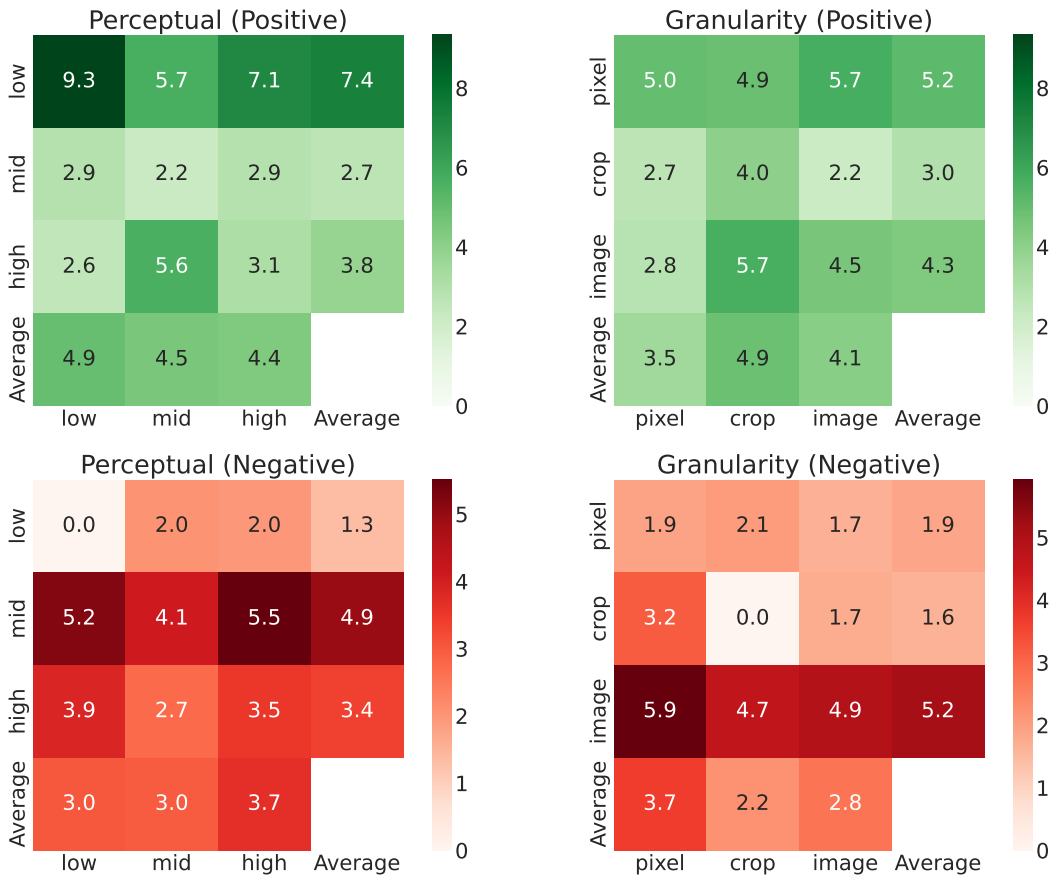


Figure A.10. Qwen2.5-VL 3B category wise heatmaps

Clique Type	Model Size	Cliques
Positive	3B	{AS, RR, VS}, {MR, RR, VS}, {RD, RR, VC}, {FC, RD, RR}, {MR, RD, RR}, {RD, SC, VC}, {FC, JG, OL}
	7B	{MR, RD, RR, VS}, {MR, RR, SR}, {AS, MR, RR}, {AS, MR, OL}, {CN, MR, OL}
	32B	{AS, CN, FC, JG, RD, RR, SR, VC, VS}, {AS, CN, FD}
Negative	3B	{AS, CN, OL, SR}, {CN, FD, OL, SR}, {CN, FD, JG, SR}, {OL, SR, VS}, {AS, FC, SR}, {AS, OL, SC}, {AS, OL, VC}, {FD, OL, VC}
	7B	{CN, FC, JG }, {FD, JG, SC }, {FD, SC, VS}, {CN, SC, VS}, {CN, JG, SC}
	32B	{FD, MR, OL, SC}

Table A.1. List of all positive and negatives cliques for all model sizes (3B, 7B, 32B) for Qwen-2.5-VL.

A.7. LoRA Weights Analysis

In this section, we analyze the cosine similarity of LoRA-finetuned weights across tasks to assess whether certain tasks induce more similar parameter updates, thereby revealing shared structure or transferable representations. For this analysis, we focus

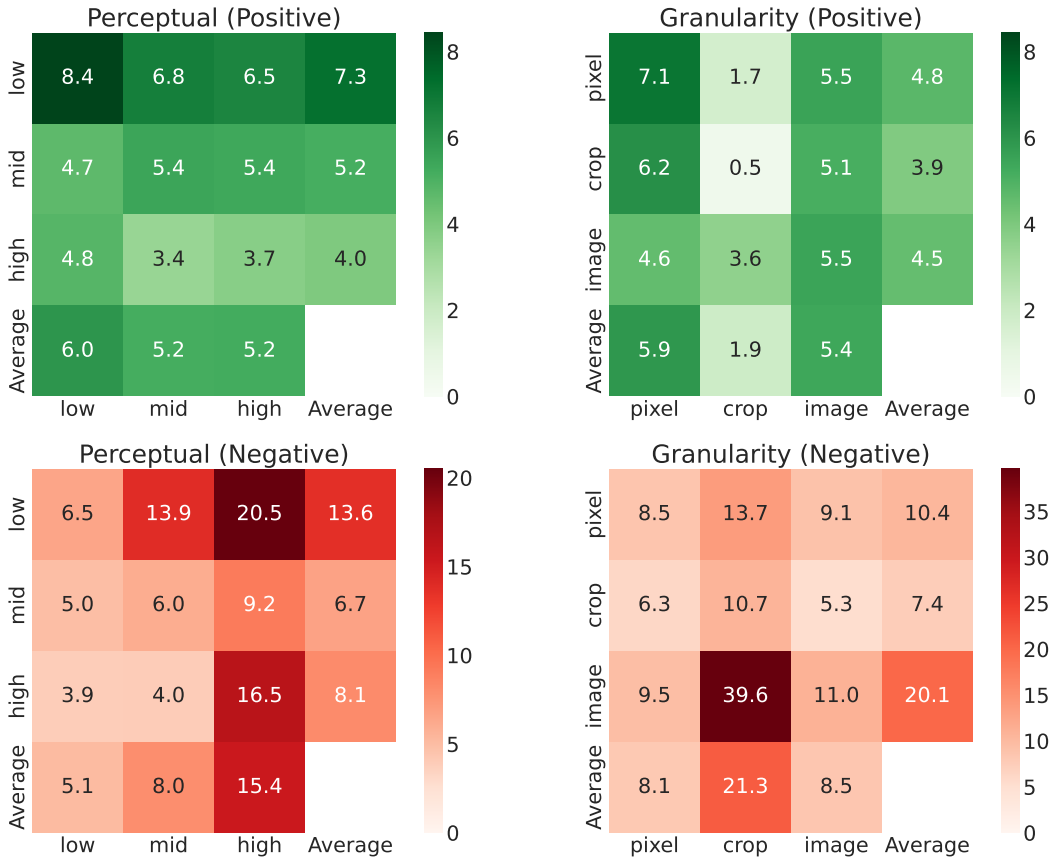


Figure A.11. Qwen2.5-VL 7B category wise heatmaps

on the output projection weights from the final layer, as they exhibited the highest variance across all the layers. Figure A.19, Figure A.20 and Figure A.21, show the resulting heatmaps for Qwen2.5-VL 3B, 7B, and 32B, respectively. Across all models, the strongest similarities appear among the Visual Similarity, Jigsaw, and Art Style tasks. We hypothesize that this arises because these are multi-image tasks, requiring comparable skills such as reasoning over pairs of images, assessing similarity, or aligning image composition. Consistent with the model-size trend, the 32B model exhibits the highest overall cosine similarity, suggesting stronger cross-task alignment in larger models. Interestingly, the 3B model shows higher similarities than the 7B model, which may be attributable to architectural differences: the 3B variant has 35 layers, whereas the 7B has 27 wider layers. A deeper interpretability analysis of these task-induced representations remains an avenue for future work.

A.8. Generalization to Other Models

We further assess whether the transfer patterns observed in Qwen2.5-VL models generalize to other VLM architectures by repeating our experiments on Llava1.5-13B. In Figure A.22, we illustrate the PGF heatmap across BLINK tasks using the Llava1.5-13B model. Qualitatively, we find that Visual Similarity, Art Style, and Jigsaw again form a coherent positive-transfer clique, aligning closely with the structure observed in Qwen2.5-VL. Likewise, Relative Depth consistently emerges as a sponge task, reinforcing its model-agnostic sensitivity to finetuning across architectures.

A.9. Task Graph Visualizations

We provide an ablation on the percentile of edges shown for visualization of the task graph in Figure A.23. We ablate on the Qwen-2.5-VL 32B model and provide visualizations for 25th, 50th, 75th and 100th percentile of edges.

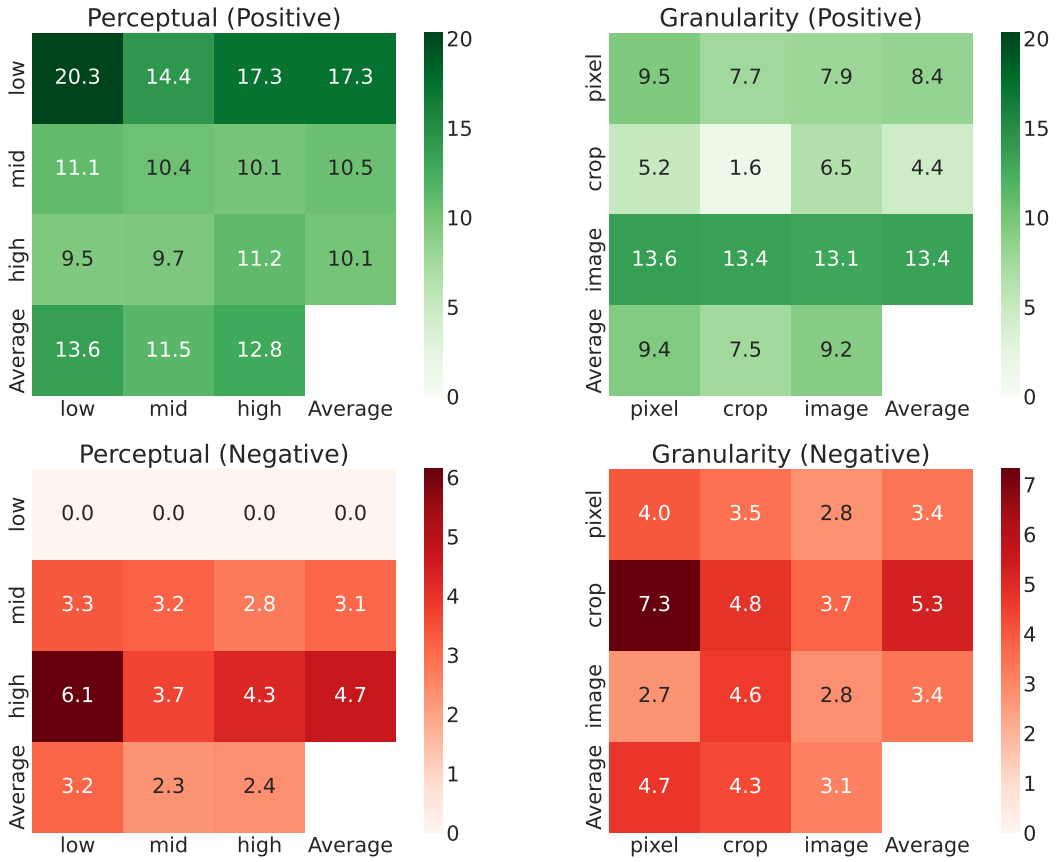
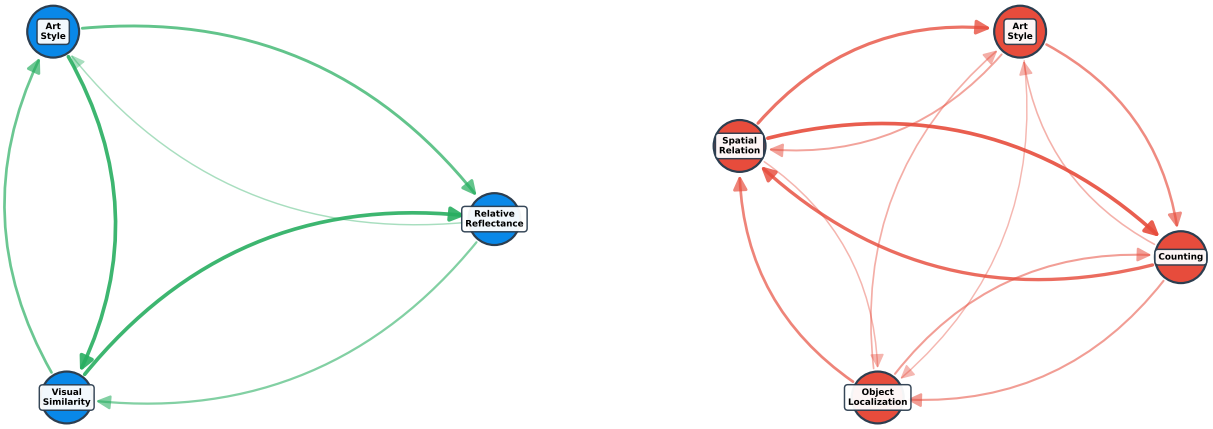


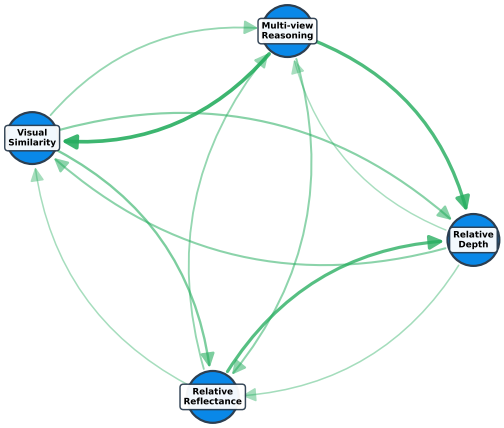
Figure A.12. Qwen2.5-VL 32B category wise heatmaps



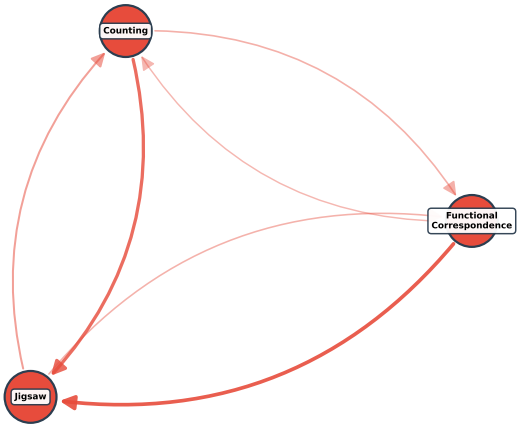
(a) Positive Clique

(b) Negative Clique

Figure A.13. Largest (a) positive and (b) negative clique for Qwen-2.5-VL 3B.

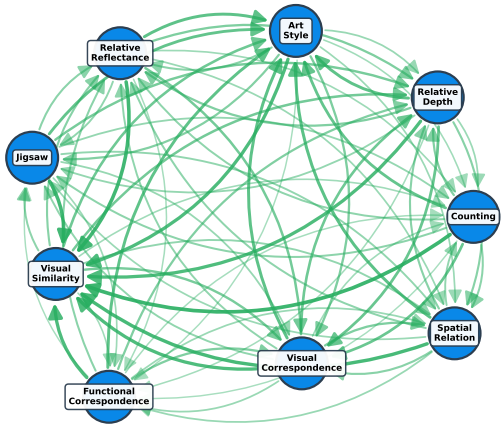


(a) Positive Clique

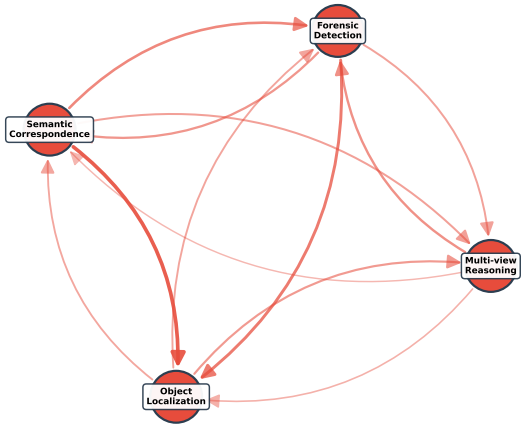


(b) Negative Clique

Figure A.14. Largest (a) positive and (b) negative clique for Qwen-2.5-VL 7B.



(a) Positive Clique



(b) Negative Clique

Figure A.15. Largest (a) positive and (b) negative clique for Qwen-2.5-VL 32B.

A.10. PGF with Best Performance Ceiling

To examine how the choice of ceiling U influences PGF, we replace the original ceiling with the best observed performance on the target task. The resulting effects are demonstrated in Figure A.24, Figure A.25, and Figure A.26. As expected, these plots exhibit a sequence of PGF scores equal to 1 along the diagonal, since direct supervision typically yields the highest performance.

A.11. Broader Impact

Vision Language Models (VLMs) are increasingly being deployed in real-world systems like robotics, surveillance, autonomous vehicles, etc. Deploying VLMs in these critical domains requires a comprehensive understanding of the impact of finetuning on various tasks. Our findings demonstrate, for the first time, how finetuning on one task impacts performance across other tasks. This may help directly help practitioners design efficient and reliable finetuning pipelines. For example, identifying

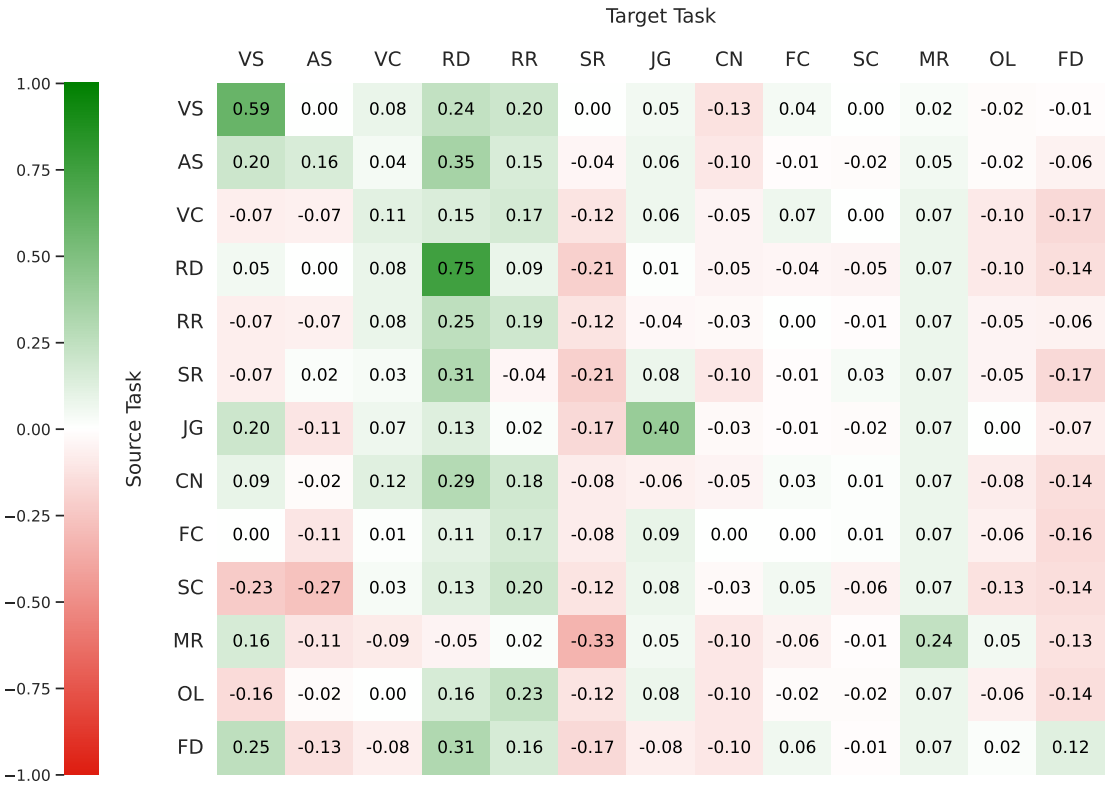


Figure A.16. PGF Heatmap for Qwen-2.5-VL 3B trained on 25% of the original training steps.

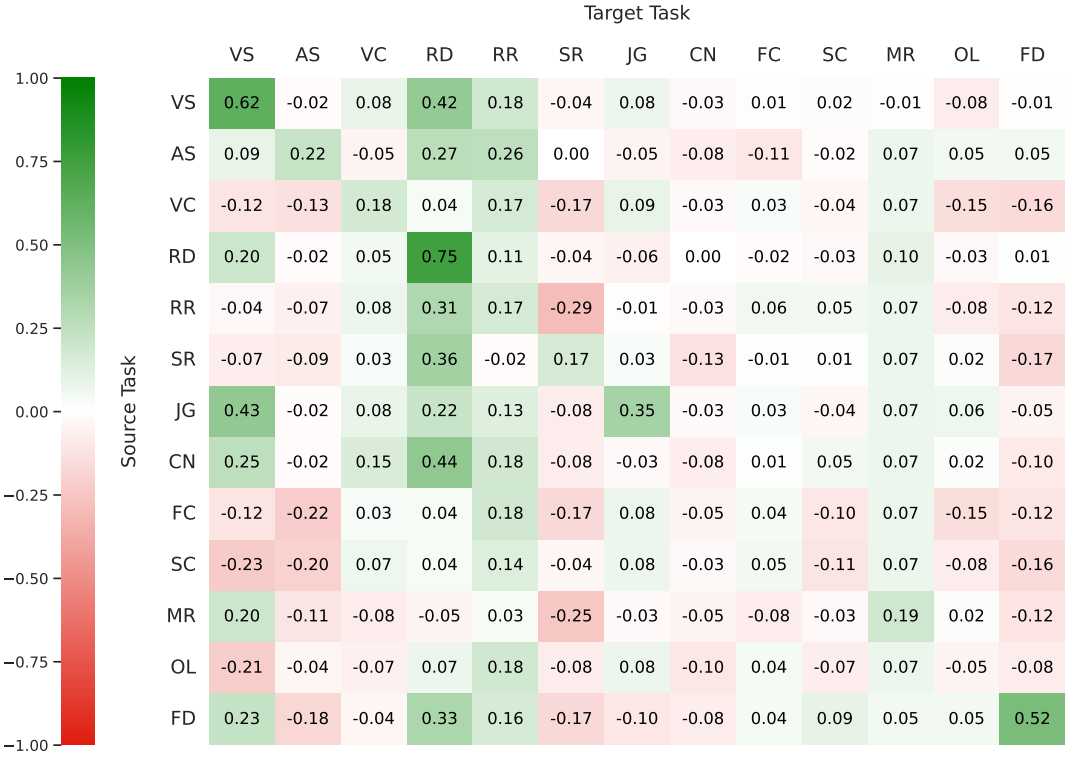


Figure A.17. PGF Heatmap for Qwen-2.5-VL 3B trained on 50% of the original training steps.

Task	Description	Source Dataset	Hyperparams
Visual Similarity	*Given a reference image alongside two alternatives, identify the image most visually similar to the reference.*	DreamSim (Nights) [5]	15,914 examples, 2500 steps, 1e-4 lr
Counting	*Given an image, a counting-related question, and 4 options, choose the correct answer.*	TallyQA [1]	250k examples, 1 epoch, 1e-4 lr
Relative Depth	*Decide which of two specified points is closer.*	Depth in the Wild + Human Annotations [4]	210k examples, 1 epoch, 1e-4 lr
Jigsaw	*Choose the image that completes the scene.*	TARA [6]	11,837 examples, 920 steps, 1e-4 lr
Art Style	*Given a reference painting and two candidate paintings, identify which shares the same art style.*	WikiArt[‡]	100k examples, 1000 steps, 1e-4 lr
Functional Correspondence	*Match a reference point in one image with the best corresponding point among 4 options in another image, based on functional affordances.*	FunKPoint [9]	100k examples, 2000 steps, 1e-4 lr
Semantic Correspondence	*Given a point in a reference image, choose the most semantically similar point among 4 options in another image.*	Spair-71k [12]	36k examples, 5 epochs, 1e-4 lr
Spatial Relation	*Identify the spatial relationship between objects in an image.*	Visual Spatial Reasoning [11]	7k examples, 5 epochs, 1e-4 lr
Object Localization	*Given an image and two bounding boxes (one ground-truth, one perturbed), choose the correct bounding box.*	LVIS [8]	18,912 examples, 1480 steps, 1e-4 lr
Visual Correspondence	*Identify the same point across two input images. One image has 1 point, the other has 4 candidate points.*	HPatches [2]	6k examples, 10 epochs, 1e-4 lr
Multi-view Reasoning	*Predict the direction of camera motion from two views.*	Wild 6D [7]	4k examples, 10 epochs, 1e-4 lr
Relative Reflectance	*Decide which of two pixels is darker, or whether they have similar reflectance.*	Intrinsic Images in the Wild + Human Annotations [3]	14k examples, 10 epochs, 1e-4 lr
Forensic Detection	*Identify synthetic images from a mixture of real and synthetic samples.*	Synthetic: COCO captions [10] + Stable Diffusion XL Real: COCO captions + Web search	60,518 examples, 100 steps, 1e-4 lr

Table A.2. Overview of tasks used in our evaluation. Each task is paired with its source dataset and finetuning setup. The number of examples, epochs/steps, and lr are specified for each task.

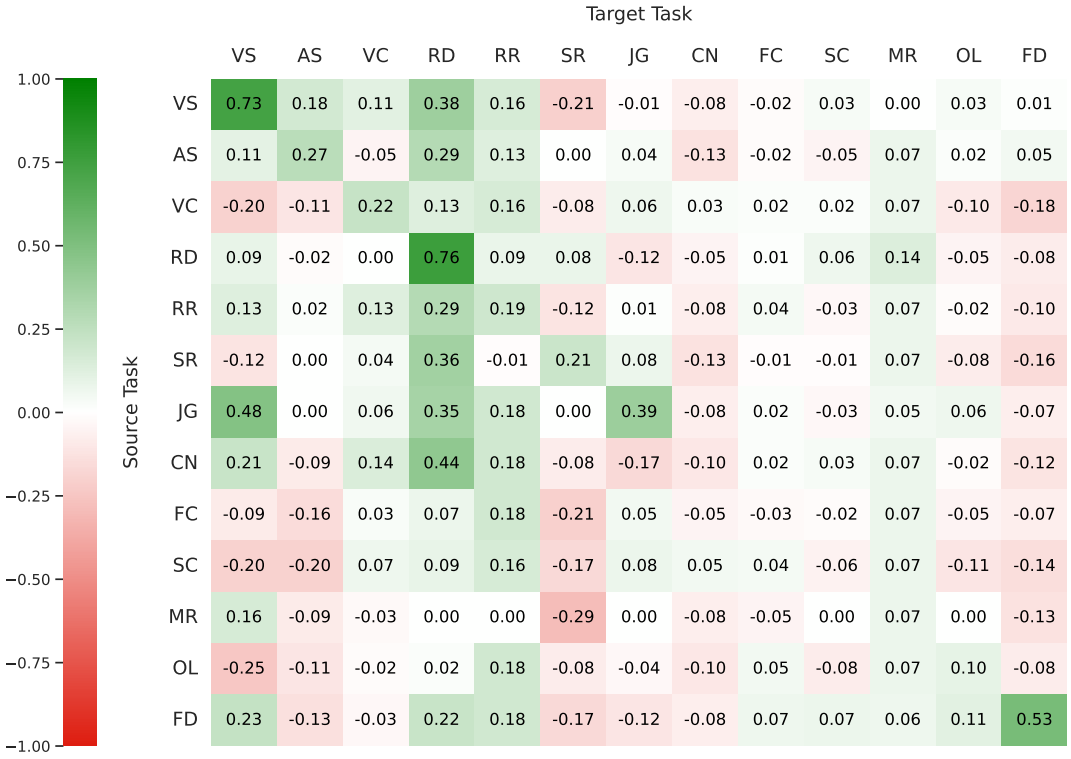


Figure A.18. PGF Heatmap for Qwen-2.5-VL 3B trained on 75% of the original training steps.

tasks that interfere with other tasks and ones that are highly transferable can reduce unexpected outcomes during deployment. Furthermore, PGF guided data selection can lower costs for users and democratize VLM finetuning. At the same time, our work highlights risks that arise from unintended transfer effects. Negative transfer between certain tasks indicates that naively finetuning VLMs for specialized capabilities can silently degrade other perception abilities, which may be consequential in safety-critical domains such as medical imaging or navigation. Although our benchmarks are standardized, real-world applications involve more diverse and noisy data distributions, where interference may be more severe. We encourage practitioners to apply transferability analyses before deploying VLMs in high-stakes settings. Overall, our analysis contributes to transparency by uncovering the patterns of task transfer. This underscrores the need for more comprehensive evaluation benchmarks for VLMs, ones that measure both performance and inter-task correlations. In future, we plan to extend this analysis to open-ended generation tasks, multiple languages, ensuring transfer behaviors generalize across diverse contexts.

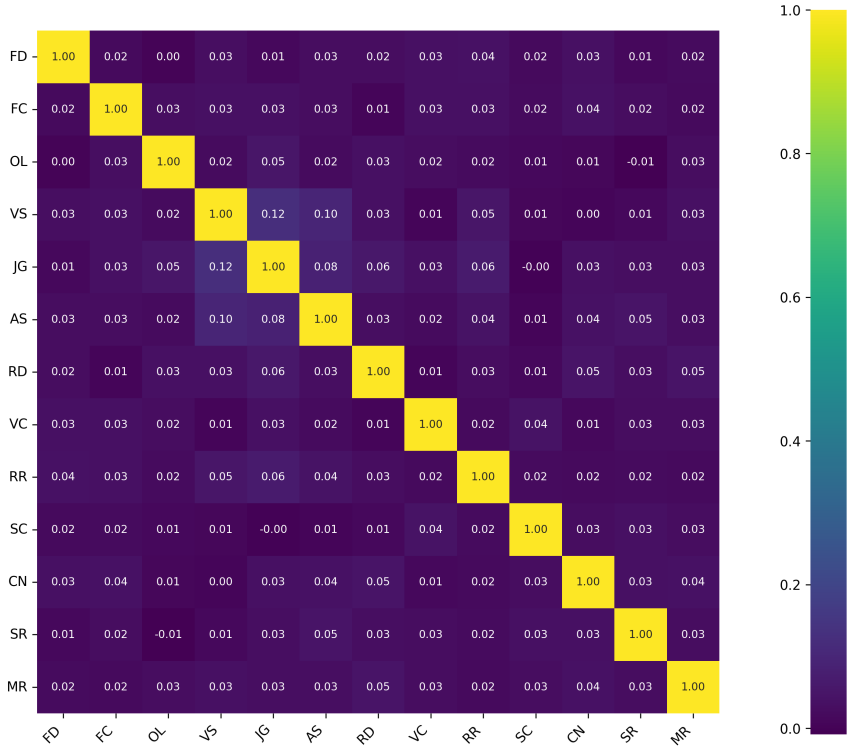


Figure A.19. Cosine Similarity of LoRA weights of the output projection from layer 35 (last layer) after finetuning Qwen2.5VL-3B.

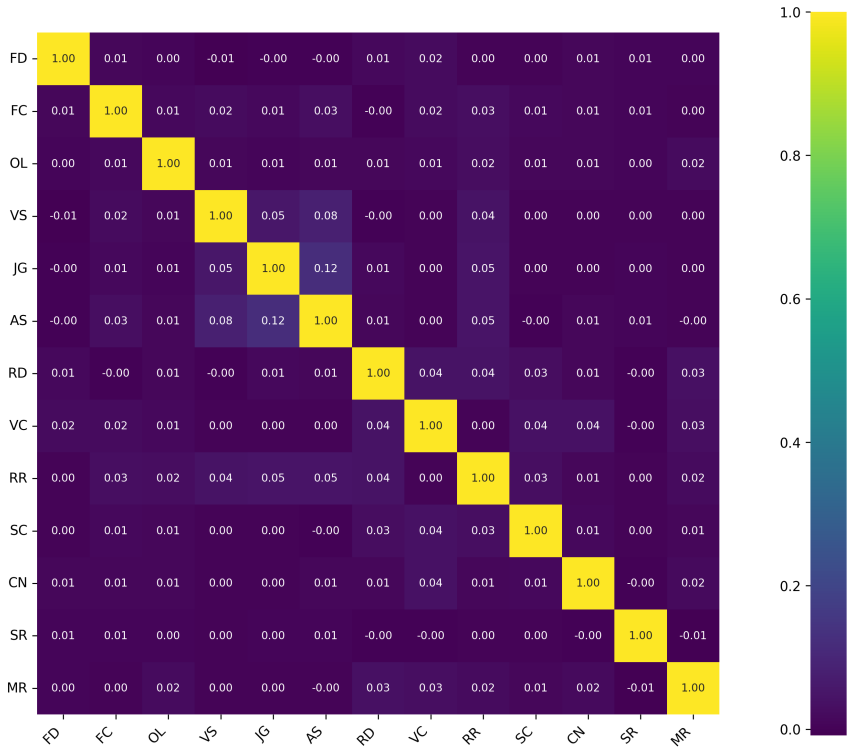


Figure A.20. Cosine Similarity of LoRA weights of the output projection from layer 27 (last layer) after finetuning Qwen2.5VL-7B.

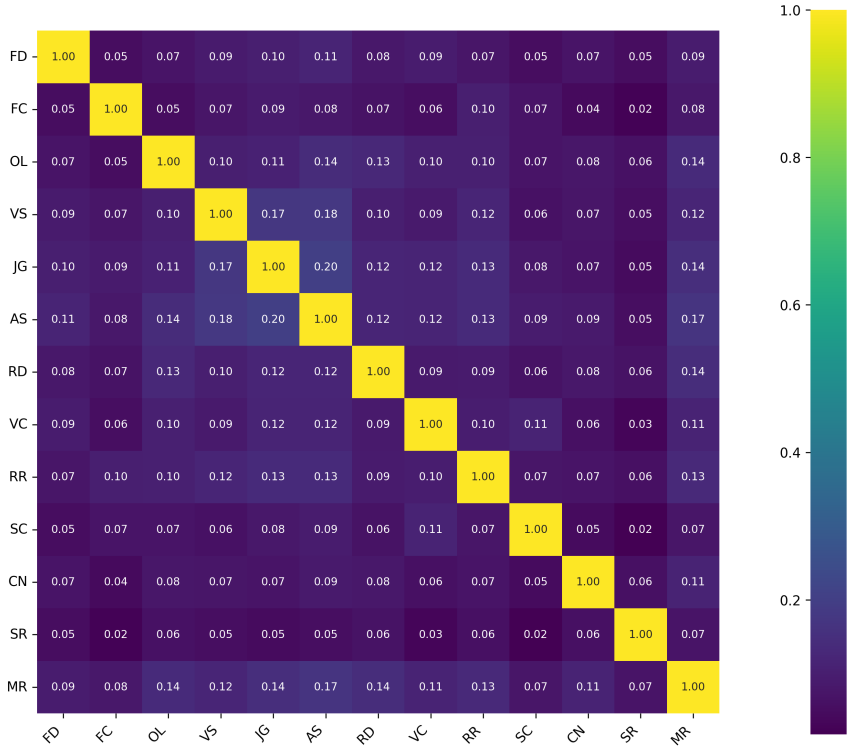


Figure A.21. Cosine Similarity of LoRA weights of the output projection from layer 65 (last layer) after finetuning Qwen2.5VL-32B.

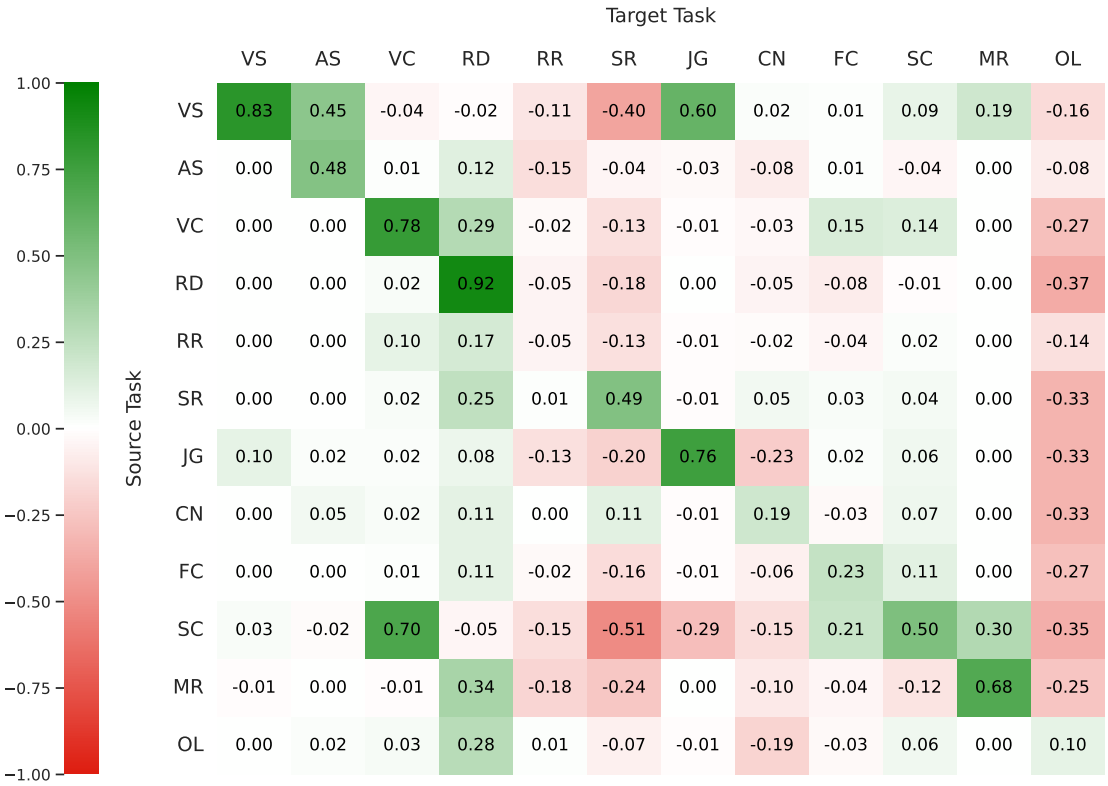
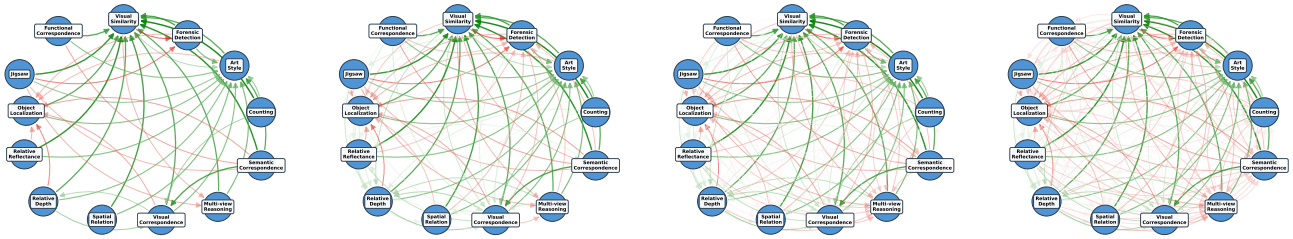


Figure A.22. PGF Heatmap for the LLaVA V1.5 13B Model.



(a) 25th percentile (b) 50th percentile (c) 75th percentile (d) 100th percentile

Figure A.23. Visualization of Qwen-2.5-VL 32B task graph with varying percentile of edges shown.

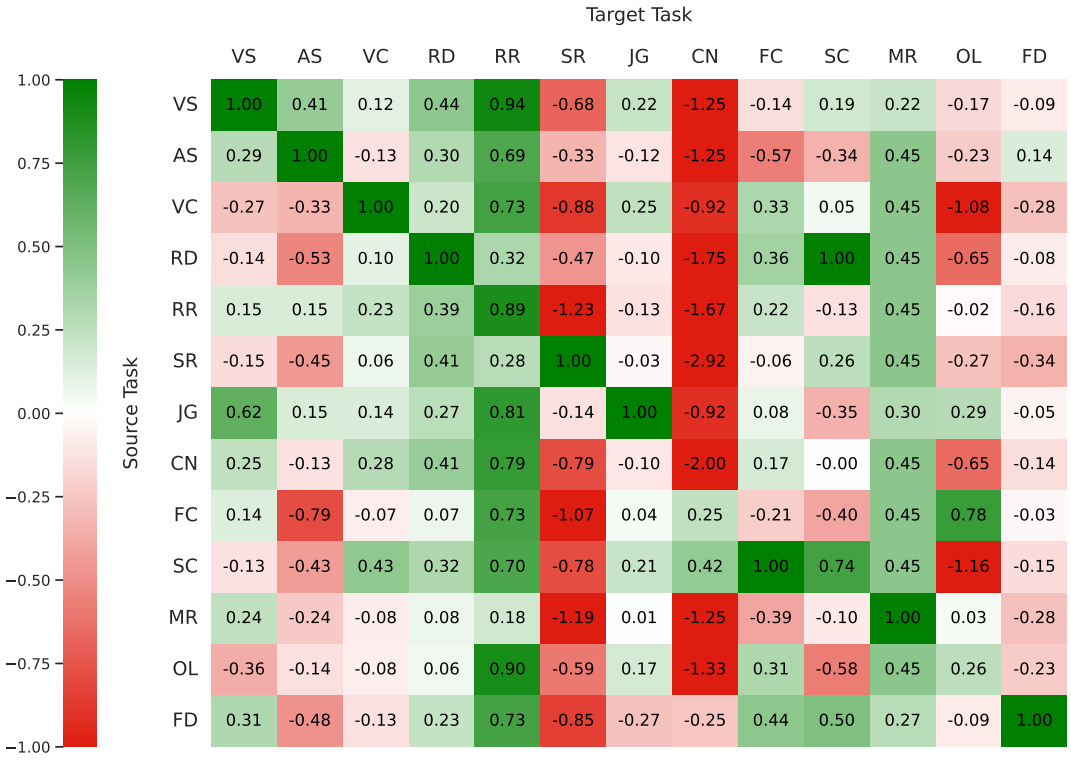


Figure A.24. Best-bound PGF Heatmap for Qwen-2.5-VL 3B.

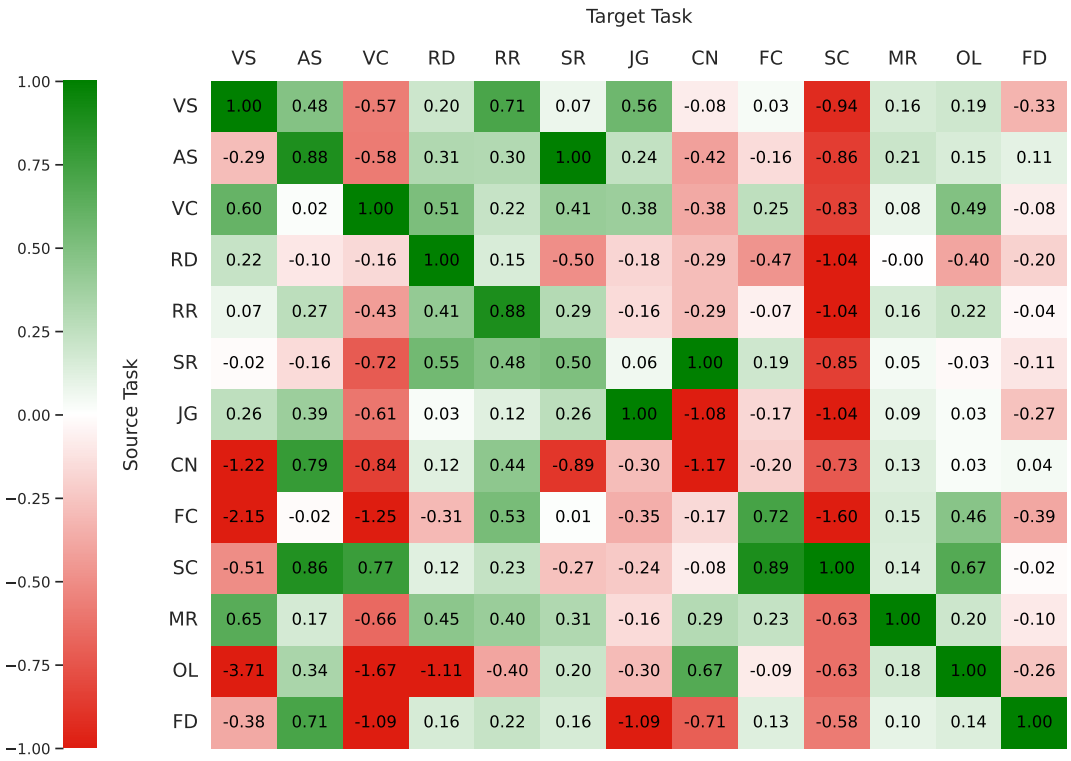


Figure A.25. Best-bound PGF Heatmap for Qwen-2.5-VL 7B.

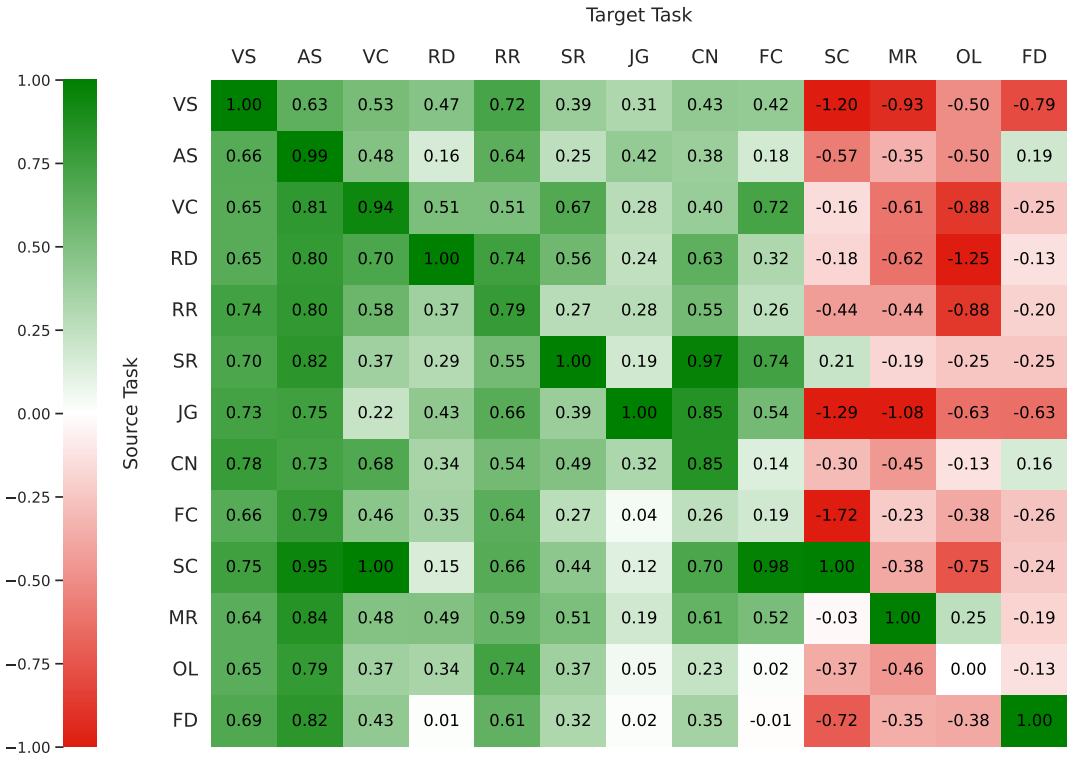


Figure A.26. Best-bound PGF Heatmap for Qwen-2.5-VL 32B.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions, 2018. 11

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017. 11

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 11

[4] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 11

[5] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, pages 50742–50768, 2023. 11

[6] Xingyu Fu, Ben Zhou, Ishaan Preetam Chandratreya, Carl Vondrick, and Dan Roth. There's a Time and Place for Reasoning Beyond the Image. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 11

[7] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. In *Advances in Neural Information Processing Systems*, 2022. 11

[8] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 11

[9] Zihang Lai, Senthil Purushwalkam, and Abhinav Gupta. The functional correspondence problem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15772–15781, 2021. 11

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 11

[11] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 11

[12] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 11

[13] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020. 5