A More Related Work

Vision-Language Models (VLMs). Vision–Language Models (VLMs) have significantly advanced in visual understanding, in part due to techniques such as Visual Instruction Tuning [23]. Subsequent work has further enhanced these models through scaling to high-resolution images using any-res [25] and windowed attention [12] supporting high fidelity tasks such as fine-grained VQA and OCR. In addition, expanding training datasets to include a diverse range of tasks—particularly multi-image tasks [18]—has been shown to improve performance on video tasks, demonstrating task transfer capabilities from multi-image to the video domain. Reasoning within VLMs has also been studied through Chain-of-Thought-style approaches, using bounding-box prediction or region selection as reasoning steps [35, 30, 5].

BLINK Benchmark. The BLINK Benchmark [10] focuses on the visual perception capabilities of VLMs. The benchmark comprises 14 tasks and each task contains multiple choice questions with single or multiple (2-5) images per question. A detailed list can be found in Table 1.

Task Transfer. The Taskonomy framework [37] investigates task transferability across a wide range of computer vision problems (e.g., image classification, semantic segmentation, depth prediction, image inpainting). Their approach involves pretraining an encoder on a source task and then training a task-specific decoder on a target task, enabling estimation of transferability scores between tasks. Sundaram et al. [32] show that finetuning vision backbone models on image similarity triplets (similar to preference tuning in LLMs) benefits a variety of downstream tasks, such as depth prediction, counting, image retrieval and segmentation. The authors argue that doing so aligns the model's latent representation with human preferences, thus leading to performance improvements. Huan et al. [15] examine how finetuning models on mathematical reasoning tasks affects their performance on both general reasoning and non-reasoning tasks. They also introduce a task-transferability index—defined as the accuracy gain relative to baseline scores—to quantify these interactions.

B Behavior of Performance Gap Factor (PGF)

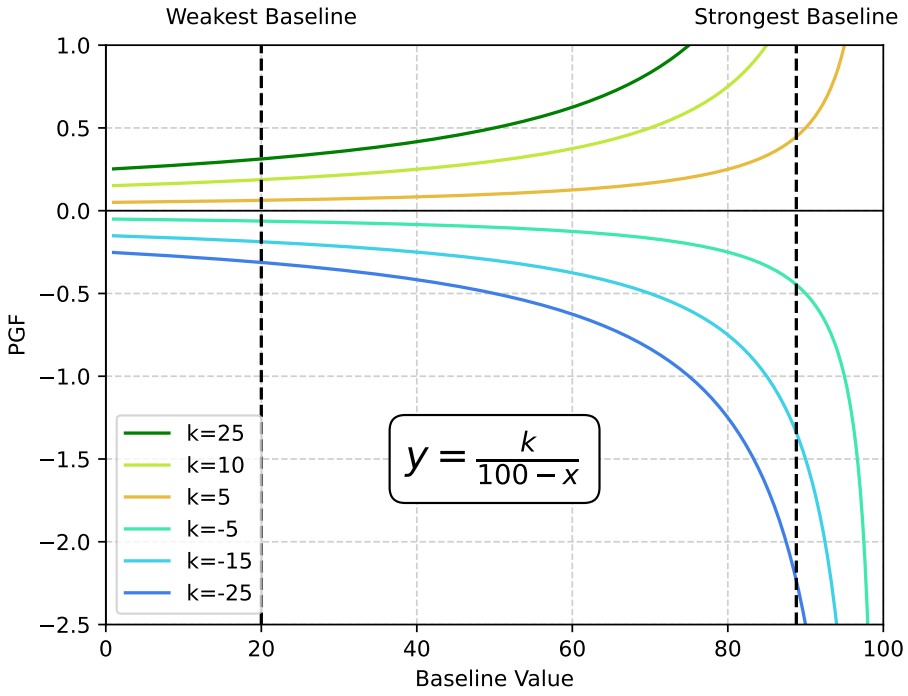


Figure 4: Behavior of PGF as a function of baseline accuracy (x) and change after finetuning (k).

Figure 4 illustrates how the Perfection Gap Factor (PGF) varies with baseline performance x and accuracy change k after finetuning. Several numerical properties emerge:

- **Positive Bound**: For improvements ($k > 0$), PGF is capped at 1, achieved when finetuning fully closes the gap to perfection ($k = 100 - x$).
- **Negative Bound**: For deterioration ($k < 0$), PGF admits a finite lower bound due to accuracy discreteness. With m evaluation questions, the highest baseline strictly below 100% is $x = 100\,(1 - \frac{1}{m})$. The worst deterioration is $k = -x$ (accuracy drops to zero), yielding

$$\text{PGF}_{\min} \;=\; \frac{-x}{100 - x} \;=\; \frac{-100(1 - \frac{1}{m})}{100/m} \;=\; -(m - 1).$$

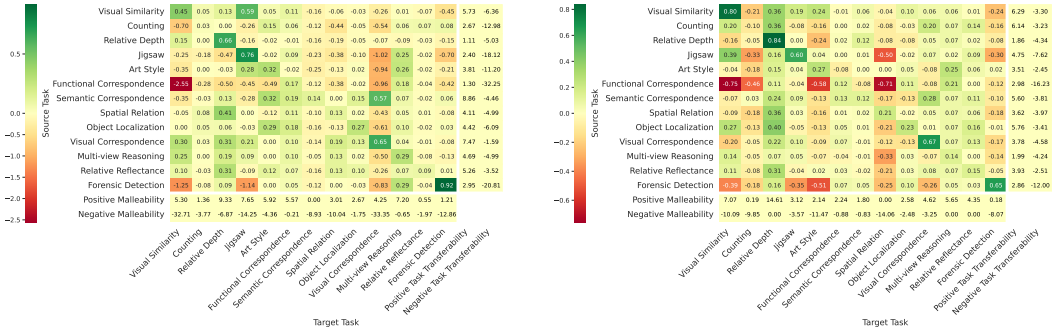
  For instance, with $m = 200$ questions, $\text{PGF}_{\min} = -199$. The worst-case deterioration therefore grows linearly with m.
- **Asymmetry**: Since positive PGF is capped at 1 but negative PGF can reach $-(m-1)$, PGF is inherently asymmetric, motivating our separate study of positive vs. negative transferability.
- **Ceiling Sensitivity**: Near-perfect baselines amplify PGF: small accuracy shifts yield disproportionately large values. This highlights ceiling-level improvements while penalizing degradations more harshly.

C Implementation Details

All training is performed on 8xA100s 40GB. DeepSpeed [29] ZeRO-2 is used for Qwen-2.5-VL 3B and 7B, while DeepSpeed [29] 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 for all tasks except Object Localization. For Object Localization, we use a rank of 64 to ensure convergence. α is set to 16 for all tasks. Task-wise training details are mentioned in Table 1.

D Additional Results

D.1 Detailed Heatmaps for Qwen2.5-VL 7B and 3B



(a) Qwen-2.5-7B (b) Qwen-2.5-3B

Figure 5: Task transferability analysis for smaller Qwen models.

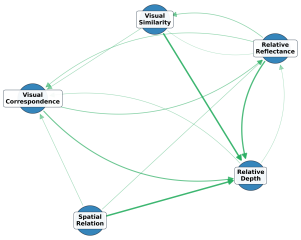
D.2 Detailed Clique Graphs

We present the 9 positive and 2 negative cliques found in Qwen-2.5-VL 32B, as well as examine similar trends in Qwen-2.5-VL 3B and 7B in Figure 2.

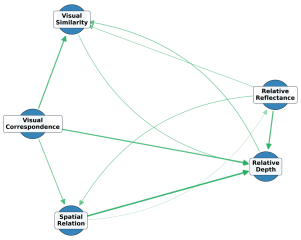
[1] https://huggingface.co/datasets/keremberke/painting-style-classification

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) [8]	15,914 examples, 10 epochs, 1e-3 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 [7]	420k examples, 1 epoch, 1e-4 lr
Jigsaw	*Choose the image that completes the scene.*	TARA [9]	11,837 examples, 5 epochs, 1e-3 lr
Art Style	*Given a reference painting and two candidate paintings, identify which shares the same art style.*	WikiArt[1]	100k examples, 500 steps, 1e-3 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 [17]	100k examples, 1000 steps, 1e-3 lr
Semantic Correspondence	*Given a point in a reference image, choose the most semantically similar point among 4 options in another image.*	Spair-71k [28]	36k examples, 5 epochs, 1e-4 lr
Spatial Relation	*Identify the spatial relationship between objects in an image.*	Visual Spatial Reasoning [21]	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 [13]	18,912 examples, 10 epochs, 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 [3]	6k examples, 10 epochs, 1e-4 lr
Multi-view Reasoning	*Predict the direction of camera motion from two views.*	Wild 6D [11]	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 [4]	14k examples, 10 epochs, 1e-4 lr
Forensic Detection	*Identify synthetic images from a mixture of real and synthetic samples.*	Synthetic: COCO captions [20] + Stable Diffusion XL Real: COCO captions + Web search	60,518 examples, 500 steps, 1e-3 lr

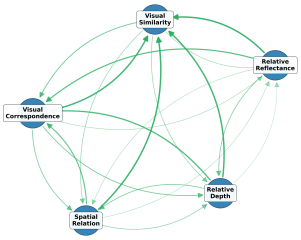
Table 1: Overview of tasks used in our evaluation. Each task is paired with its source dataset and fine-tuning setup. The number of examples, epochs/steps, and lr are specified for each task.



(a) Qwen-2.5-VL 3B

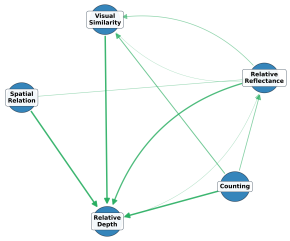


(b) Qwen-2.5-VL 7B

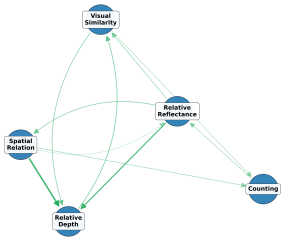


(c) Qwen-2.5-VL 32B

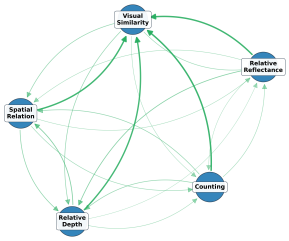
(i) Positive Clique 1



(a) Qwen-2.5-VL 3B

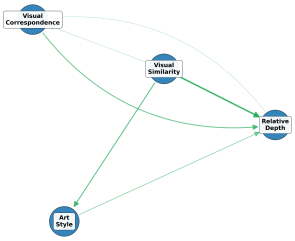


(b) Qwen-2.5-VL 7B

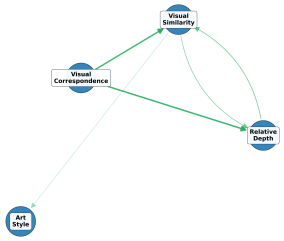


(c) Qwen-2.5-VL 32B

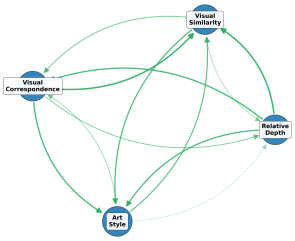
(ii) Positive Clique 2



(a) Qwen-2.5-VL 3B

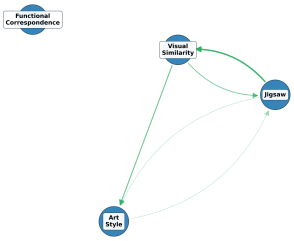


(b) Qwen-2.5-VL 7B

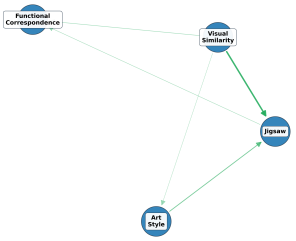


(c) Qwen-2.5-VL 32B

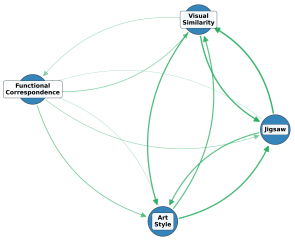
(iii) Positive Clique 3



(a) Qwen-2.5-VL 3B

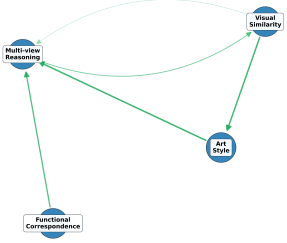


(b) Qwen-2.5-VL 7B

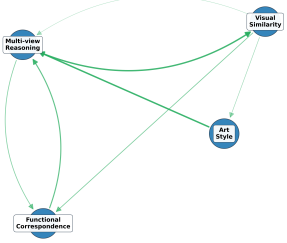


(c) Qwen-2.5-VL 32B

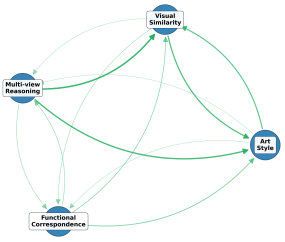
(iv) Positive Clique 4



(a) Qwen-2.5-VL 3B

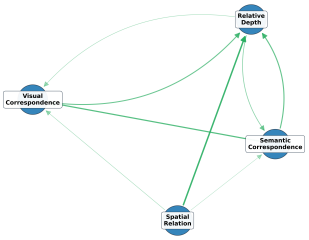


(b) Qwen-2.5-VL 7B

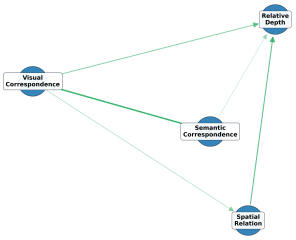


(c) Qwen-2.5-VL 32B

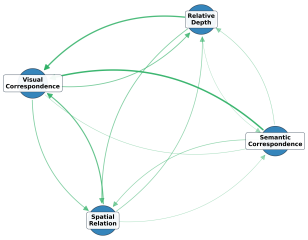
(v) Positive Clique 5



(a) Qwen-2.5-VL 3B

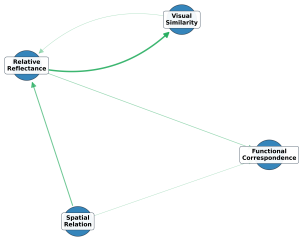


(b) Qwen-2.5-VL 7B

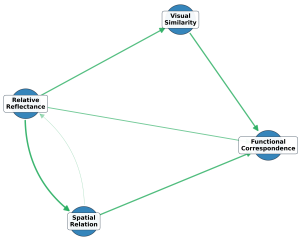


(c) Qwen-2.5-VL 32B

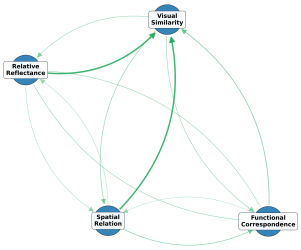
(vi) Positive Clique 6



(a) Qwen-2.5-VL 3B

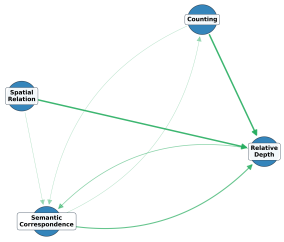


(b) Qwen-2.5-VL 7B

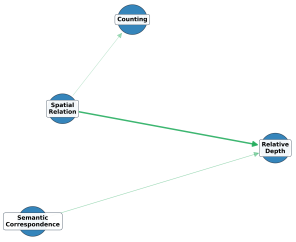


(c) Qwen-2.5-VL 32B

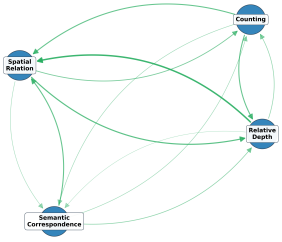
(vii) Positive Clique 7



(a) Qwen-2.5-VL 3B

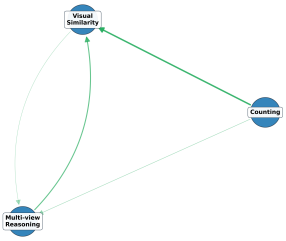


(b) Qwen-2.5-VL 7B

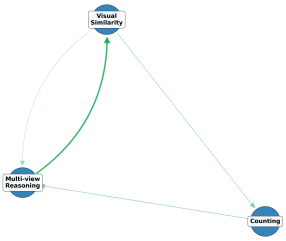


(c) Qwen-2.5-VL 32B

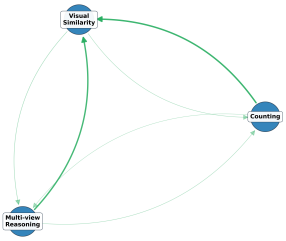
(viii) Positive Clique 8



(a) Qwen-2.5-VL 3B

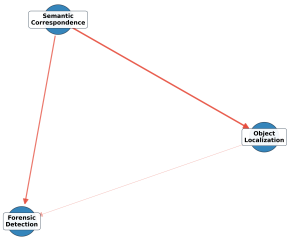


(b) Qwen-2.5-VL 7B

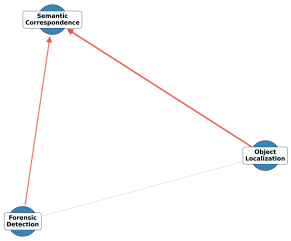


(c) Qwen-2.5-VL 32B

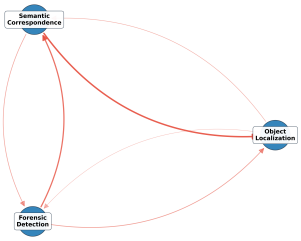
(ix) Positive Clique 9



(a) Qwen-2.5-VL 3B

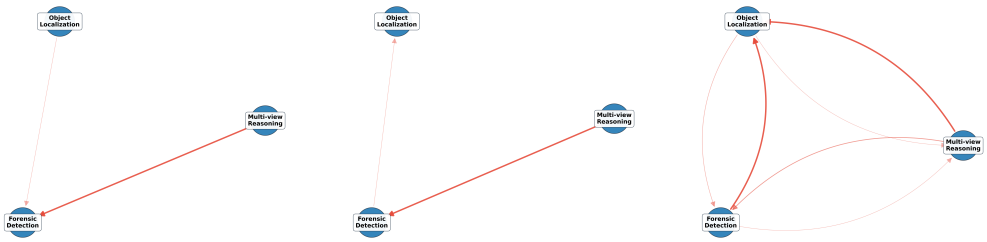


(b) Qwen-2.5-VL 7B



(c) Qwen-2.5-VL 32B

(x) Negative Clique 1



(a) Qwen-2.5-VL 3B (b) Qwen-2.5-VL 7B (c) Qwen-2.5-VL 32B
(xi) Negative Clique 2

Table 2: Cliques across all model sizes