

# Understanding Task Transfer in Vision-Language Models

Bhuvan Sachdeva\* Karan Uppal\* Abhinav Java\* Vineeth N. Balasubramanian  
Microsoft Research India

Project page: <https://aka.ms/task-transfer-vlms>

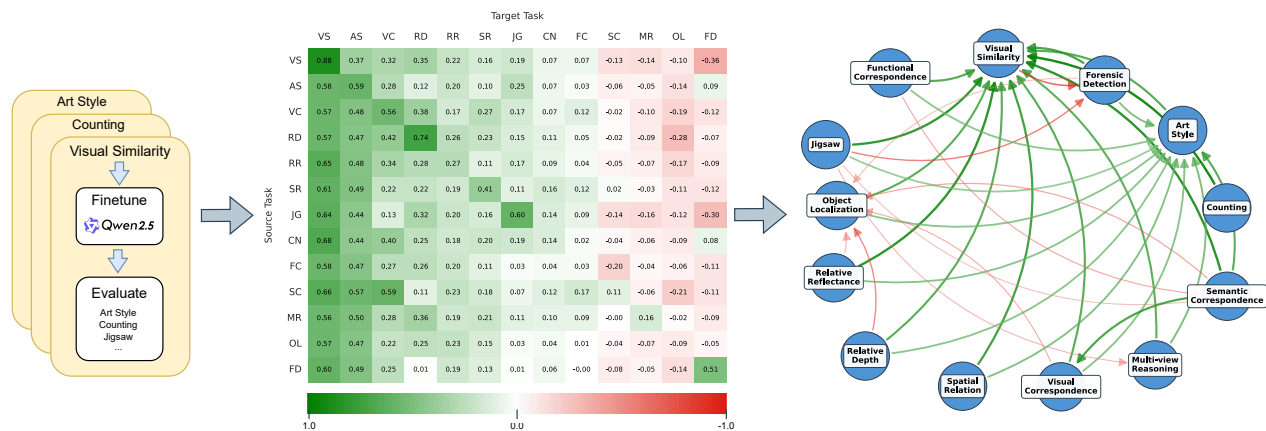


Figure 1. One finetune, many fates: Finetuning Qwen-2.5-VL 32B on perception tasks creates a structured map of transfer capabilities. (The list of perception tasks considered can be found in Table 2.)

## Abstract

*Vision-Language Models (VLMs) perform well on multi-modal benchmarks but lag behind humans and specialized models on visual perception tasks like depth estimation or object counting. Finetuning on one task can unpredictably affect performance on others, making task-specific finetuning challenging. In this paper, we address this challenge through a systematic study of task transferability. We examine how finetuning a VLM on one perception task affects its zero-shot performance on others. We introduce Perfection Gap Factor (PGF), a normalized metric that measures change in performance as a result of task transfer. We utilize PGF to compute Task Transferability, which captures both the breadth and the magnitude of transfer induced by a source task. Using three open-weight VLMs evaluated across 13 perception tasks, we construct a task transfer graph that reveals previously unobserved relationships among perception tasks. Our analysis uncovers patterns of positive and negative transfer, identifies groups of tasks that mutually influence each other, organizes tasks into personas based on their transfer behavior and demonstrates how PGF can guide data selection for more efficient training. These findings highlight both opportunities for positive transfer and risks of negative interference, offering actionable guidance for advancing VLMs.*

## 1. Introduction

Vision Language Models [1, 10, 11, 13, 14, 24] have demonstrated significant progress in understanding visual information in recent years, as reflected in their performance across well-known benchmarks such as MMMU [27], DocVQA [18], InfoVQA [17], and TextVQA [21]. Visual instruction tuning [13] has helped adapt Large Language Models (LLMs) to parse visual input by finetuning a small number of parameters to align a visual encoder (e.g., CLIP [19]) with a given LLM backbone. Despite this progress, careful analysis has shown that VLMs fall short in many visual understanding tasks, most often due to their limitations in visual perception [5, 23]. Understanding the limits of VLMs on visual perception tasks, especially ones that are natural to humans and serve as building blocks that scaffold on to more complex visual tasks remains an urgent need, in order to provide foundational solutions to robust visual processing.

VLMs lag behind humans and specialist models on basic perception tasks such as depth estimation, object detection, and counting. For example, on the BLINK [5] leaderboard, the top performing models (GPT-4o at 60.04% and GPT-4V at 51.14%) achieve modest average performance compared to humans (95%). This has motivated practitioners

\*Equal contribution. Corresponding Author: vineeth.nb@microsoft.com

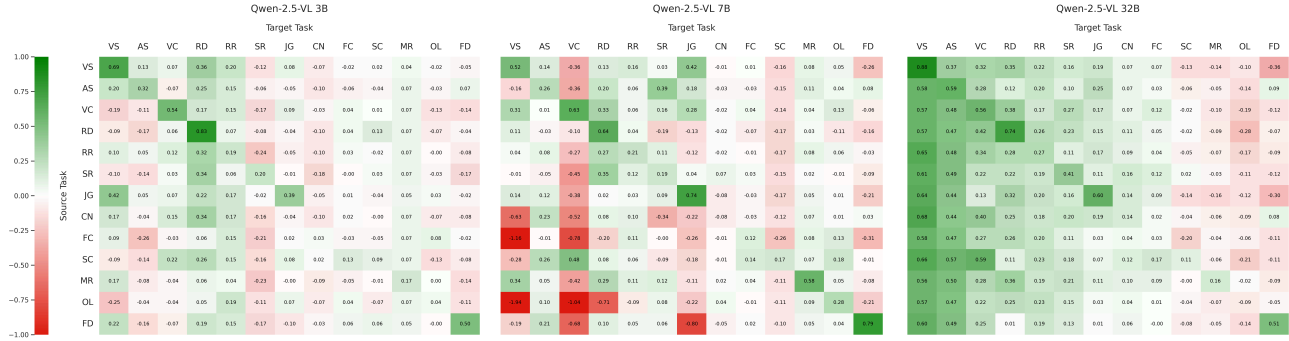


Figure 2. PGF Heatmaps for Qwen-2.5-VL model family (3B, 7B, 32B).

to finetune VLMs using parameter-efficient methods such as LoRA [8] on curated, task-specific datasets to improve performance on perception tasks.

However, little is known about how such finetuning affects a model’s other perception capabilities, particularly in modern foundation models used predominantly in zero-shot settings. Understanding this phenomenon is essential for both improving robustness and characterizing how VLMs adapt and generalize. In this work, we address this gap by investigating the following question, which to our knowledge has not been previously studied:

**How does finetuning a VLM on one perception task affect its zero-shot performance on other perception tasks?**

Prior work on task relationships in vision has largely focused on transfer learning involving finetuning on both source and target tasks [3, 28]. Other efforts have explored pretraining strategies and their downstream effects [20, 22]. In contrast, we study zero-shot cross-task transfer in VLMs: how a single-task finetuning intervention reshapes performance across a diverse set of perception tasks.

To quantify these effects, we introduce the notion of *transferability* and *malleability*, where transferability captures the effect a source task induces on other tasks through finetuning, and malleability captures how susceptible a target task is to being affected by finetuning on other tasks. These metrics encompass both the **breadth** (how many tasks are influenced) and the **magnitude** (the average strength of that influence). The metrics, *transferability* and *malleability*, are computed using *Perfection Gap Factor*, which is a normalized measure of the extent to which finetuning on a source task improves or degrades zero-shot performance on a target task. Our analysis reveals several properties of tasks, including scale-dependent sharpening of transfer, distinct behaviors across perception granularity levels, and natural clustering among mutually beneficial tasks. We further extend our study to video models and observe similar trends across model sizes. Finally, we demonstrate that PGF can guide principled selec-

tion of training data subsets to improve finetuning efficiency and reduce performance regressions.

**Key Contributions.** Our contributions are as follows:

- **Systematic Study:** We are the first to analyze how finetuning on one visual perception task affects zero-shot performance on a broad suite of other perception tasks.
- **Perfection Gap Factor:** PGF provides a transfer measure that normalizes for heterogeneous task difficulties and model baselines.
- **Task Properties:** We uncover consistent structural properties of transfer, including scale-dependent sharpening, task type dependent transfer (granularity and perceptual), and the emergence of mutually beneficial task clusters.
- **Evaluation beyond images:** We evaluate cross-task transfer in multimodal models trained on video, showing that our key findings generalize to the temporal domain.
- **Downstream use of PGF:** We demonstrate that PGF can be used to identify beneficial source tasks and construct data subsets that improve finetuning efficiency while mitigating negative transfer.

**2. Related Work**

**Benchmarks for Visual Perception.** Several benchmarks [6, 12, 17, 18] have been introduced to evaluate the progress of VLMs on visual understanding. Notably, MMMU [27] is a large-scale benchmark assessing model capabilities across 30 subjects spanning technology, engineering, art, medicine, and more. Despite its broad coverage, MMMU largely focuses on examination-style question answering and lacks core perception tasks. Other benchmarks, such as DocVQA [18] and InfoGraphicsVQA [17], combine OCR capabilities with visual understanding, while ChartQA [16] evaluates a model’s ability to parse complex charts and draw meaningful inferences. Visual Commonsense Reasoning [29] presents a challenging task in which models must reason about the intention or consequences of actions depicted in a scene. MathVista [15] is a multiple-choice mathematics dataset that requires interpreting figures or di-

agrams to answer questions, and NLVR [25] tests a broad range of linguistic phenomena through image captioning tasks. While each of these datasets highlights specific aspects of multimodal understanding, most do not explicitly measure the visual perception capabilities that humans perform naturally. BLINK [5], in contrast, aggregates over 14 datasets spanning diverse tasks and serves as a central benchmark for our experiments.

**Transferability Studies.** The Taskonomy framework [28] first introduced a framework for modeling the structure of computer vision tasks through transfer learning. Zamir et al. [28] pretrain an encoder on a source task and then perform transfer learning on a task-specific decoder to estimate transferability between tasks. However, this study is restricted to the pre foundation model era and is primarily conducted using CNN-based vision encoders and small decoders. [2] introduce a new metric to estimate the performance of transfer-learned representations from source to target task. Unlike prior works that investigate transfer-learning, we focus our study on understanding the *zero-shot task transfer* in VLMs providing novel ways to quantify it and provide actionable *finetuning insights*. Shariatnia et al. [20] compare various pretraining techniques by evaluating their zero-shot capabilities. In contrast, our work focuses on task finetuning, rather than pretraining strategies, to study how perception tasks transfer in modern foundation models. Closer to our work, [22] conduct experiments with VLMs from an evaluation perspective. Though similar, [22]’s study encompasses tasks like OCR, VQA, Captioning, Visual Reasoning, etc, on transfer learning providing insights about common biases such as length of output. [3] study the impact of various factors like dataset size, pretraining strategy on transfer in vision language models. In Natural Language Processing, [9] 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, that is the accuracy gain relative to baseline scores. We present the details of our metric in Section 3, which is different from standard accuracy gain based metrics that do not account for variance in task difficulty.

**Finetuning VLMs.** Adapting VLMs to unseen domains and tasks is an active area of research. A variety of strategies have been proposed to improve the efficiency and effectiveness of finetuning, particularly for large-scale models where full-parameter updates are computationally expensive. Parameter-efficient finetuning methods such as LoRA [8] and related techniques [7] have been widely adopted, allowing small subsets of model parameters to be updated while keeping the majority of the network frozen. These approaches have been shown to maintain or even improve downstream performance across a range of tasks, making them especially attractive for finetuning. QLoRA [4] back-propagates gradients through a frozen, 4-bit quantized model

into Low Rank adapters, making the finetuning more efficient while preserving 16-bit performance. Despite these advancements, systematic studies on how task-specific finetuning impacts the transferability of perception capabilities in VLMs remain limited, motivating our work.

### 3. Problem Formulation and Metric

We present our framework for characterizing the behavior of Vision-Language Models (VLMs) across diverse perception tasks. Our goal is to understand how finetuning on one task influences performance on others. We begin by discussing preliminaries like notation and problem setup, followed by quantifying task transferability using our proposed metric **Perfection Gap Factor (PGF)**, which provides a robust way to account for differences in task difficulty and performance ceilings.

**Preliminaries.** We consider the setting where a VLM  $\mathcal{M}$  is finetuned on a source task  $T_S$  using a source dataset  $\mathcal{D}_S^{\text{train}}$  and subsequently evaluated on a set of  $N$  target tasks  $\{T_j\}_{j=1}^N$  using target datasets  $\{\mathcal{D}_j^{\text{eval}}\}_{j=1}^N$ . We represent a VLM  $\mathcal{M}$  finetuned on a dataset  $\mathcal{D}_i$  for task  $T_i$  as  $\mathcal{M}(T_i)$ . The central question we study is on how finetuning on a task  $T_S$  affects zero-shot performance on tasks  $\{T_j\}_{j=1}^N$ , and how one can quantify such inter-task relationships. We begin by formally defining task transferability.

**Definition 1 (Task Transferability)** Let  $\mu_{i \rightarrow j}$  denote a metric that captures change in performance on a target task  $T_j$  as a result of finetuning on source task  $T_i$ . Define  $p = |\{j : \mu_{i \rightarrow j} > 0\}|$  as the number of positive scores, and  $n = |\{j : \mu_{i \rightarrow j} < 0\}|$  as the number of negative scores. The **positive** and **negative** task transferability of  $T_i$  to a set of target tasks  $\{T_j\}_{j=1}^N$  are given by:

$$\begin{aligned} \Delta(i)^+ &= \left(\frac{1-e^{-\frac{p}{N}}}{p}\right) \sum_{j=1}^N \mu_{i \rightarrow j} \mathbf{1}_{\{\mu_{i \rightarrow j} > 0\}}, \\ \Delta(i)^- &= \left(\frac{1-e^{-\frac{n}{N}}}{n}\right) \sum_{j=1}^N \mu_{i \rightarrow j} \mathbf{1}_{\{\mu_{i \rightarrow j} < 0\}}. \end{aligned} \tag{1}$$

where  $\Delta(i)^+$  and  $\Delta(i)^-$  denote positive and negative task transferability, respectively.

$\Delta(i)^\pm$  captures both the *magnitude* and the *breadth* of influence of a source task  $T_i$ . The summation term helps measure the average performance gain or degradation across the affected tasks. The exponential weighting adjusts for the number of tasks, penalizing cases where positive or negative effects occur only on a small fraction of tasks. In other words, a task with large but isolated transfer effects will be scored lower than one that provides consistent improvements across many targets. However, this metric only describes the behavior of a source task on other tasks.

To characterize how sensitive a target task is to finetuning on other tasks, we introduce the notion of malleability. A target task can be considered highly malleable if finetuning on different source tasks leads to significant change (positive or negative) in the performance on that task. To quantify this value, we aggregate the positive and negative PGF scores induced on that task by other source tasks.

**Definition 2 (Malleability)** Let  $\mu_{i \rightarrow j}$  denote a metric that captures change in performance on a target task  $T_j$  as a result of finetuning on source task  $T_i$ . Define  $p = |\{i : \mu_{i \rightarrow j} > 0\}|$  as the number of positive scores, and  $n = |\{i : \mu_{i \rightarrow j} < 0\}|$  as the number of negative scores. The **positive** and **negative** malleability of  $T_j$  to a set of source tasks  $\{T_i\}_{i=1}^N$  are given by:

$$\Theta(j)^+ = \left(\frac{1-e^{-\frac{p}{N}}}{p}\right) \sum_{i=1}^N \mu_{i \rightarrow j} \mathbf{1}_{\{\mu_{i \rightarrow j} > 0\}},$$

$$\Theta(j)^- = \left(\frac{1-e^{-\frac{n}{N}}}{n}\right) \sum_{i=1}^N \mu_{i \rightarrow j} \mathbf{1}_{\{\mu_{i \rightarrow j} < 0\}}.$$
(2)

where  $\Theta(j)^+$  and  $\Theta(j)^-$  denote positive and negative malleability, respectively.

Note: These definitions are agnostic to the choice of  $\mu_{i \rightarrow j}$ . For our analysis, we use *Perfection Gap Factor* (defined below) as the default choice, and include transferability analysis using relative gain in the supplementary.

**Perfection Gap Factor.** A central challenge in quantifying task transferability and malleability is designing a metric that is comparable across tasks with very different difficulty levels and performance ceilings. Reporting accuracy gains after finetuning can be misleading. For example, a +2% improvement on a task where the model is already near the ceiling is much more significant than the same +2% improvement on a task where the model starts very low. Conversely, small drops in accuracy near the floor may not reflect meaningful transfer failure. To address this, we introduce the *Perfection Gap Factor* (PGF), which measures how much of the remaining gap to the ceiling is closed (or opened) by finetuning on a source task. Mathematically, we define the **PGF** between a source task  $T_i$  and a target task  $T_j$  as the *ratio of performance gain to the performance gap*, i.e.,

$$\mu_{i \rightarrow j} = \frac{\text{Acc}(\mathcal{M}(T_i), T_j) - \text{Acc}(\mathcal{M}, T_j)}{U_j - \text{Acc}(\mathcal{M}, T_j) + \epsilon} \quad (3)$$

where  $\text{Acc}(\mathcal{M}, T_j)$  represents the accuracy of model  $\mathcal{M}$  on task  $T_j$ ,  $U_j$  is the upper-bound (ceiling) performance of the target task and  $\epsilon = 10^{-6}$  is added to the denominator for numerical stability. By normalizing the gain relative to the remaining gap, PGF becomes both bounded and interpretable, making it comparable across tasks. Values above zero indicate positive transfer, while negative values indicate degradation. Intuitively, PGF incorporates the following questions:

Task	Baseline (%)	After FT (%)	Ceiling (%)	Raw Gain	PGF
A	90	93	95	+3	0.60
B	40	50	95	+10	0.18
C	98	97	99	-1	-0.50

Table 1. Illustration of the Perfection Gap Factor (PGF) across three target tasks.

Task Name	Abbreviation	Perceptual Level	Granularity
Art style	AS	Mid-level	Image-level
Counting	CN	High-level	Image-level
Forensics detection	FD	High-level	Image-level
Functional corr.	FC	High-level	Pixel-level
Jigsaw	JG	Mid-level	Crop-level
Multi-view reasoning	MR	Mid-level	Image-level
Object localization	OL	High-level	Crop-level
Relative depth	RD	Low-level	Pixel-level
Relative reflectance	RR	Low-level	Pixel-level
Semantic corr.	SC	High-level	Pixel-level
Spatial reasoning	SR	Mid-level	Image-level
Visual corr.	VC	Low-level	Pixel-level
Visual similarity	VS	High-level	Image-level

Table 2. BLINK tasks with abbreviation and classification by Perceptual Level and Granularity.

- ① How much does finetuning on a source task improve the target task?
- ② What is the model’s zero-shot performance on the target task before finetuning?
- ③ What is the ceiling performance of the target task?

We illustrate PGF with the help of a toy example and later discuss its properties.

**Toy Example.** Consider three target tasks with different baselines and ceiling performance, as shown in Table 1. Although Task B shows the largest raw gain (+10), it closes only 18% of its remaining gap to perfection. In contrast, Task A, despite a smaller +3 gain, closes 60% of its available headroom. Task C illustrates the opposite case: a small drop from 98% to 97% yields a PGF of  $-0.5$ , reflecting a complete loss of the narrow headroom. This example illustrates how PGF provides a normalized and interpretable view of task transferability, enabling comparison across tasks with varying difficulty levels and performance ceilings.

## 4. Results and Analysis

**Experimental Setup.** We consider a diverse set of 13 multi-modal perception tasks<sup>1</sup>, from the widely followed BLINK Benchmark [5], listed in Table 2. A detailed description of these tasks can be found in the supplementary material. We employ three variants (3B, 7B, and 32B) from the open-weight Qwen-2.5-VL lineup [24] as base models for our experiments. These models are finetuned independently on

<sup>1</sup>We exclude the “IQ Test” task from our analysis because it was manually constructed and does not have a corresponding training set.

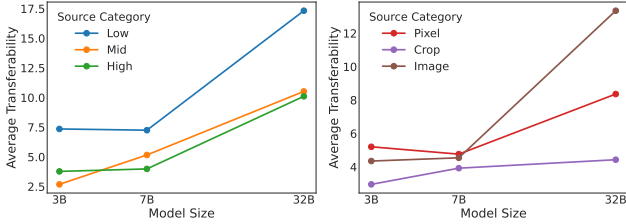


Figure 3. Average positive transferability trends across granular and perceptual levels. We observe that positive transferability increases with model size and generally low-level and image-level are highly transferable. Detailed category-wise heatmaps are provided in the supplementary material.

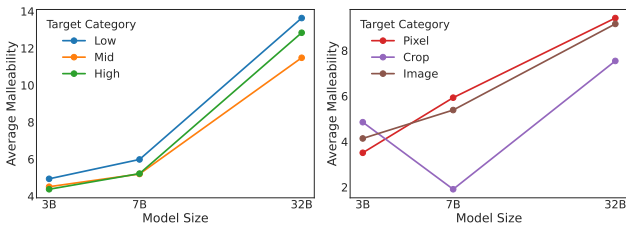


Figure 4. Average positive malleability trends across granular and perceptual levels. We observe that positive malleability increases with model size and generally low-level benefit the most from finetuning. Detailed category-wise heatmaps are provided in the supplementary material.

each task using LoRA [8], and evaluation is carried out on the validation splits of all tasks. Since BLINK itself only provides validation and test splits, we construct training data by retrieving the original datasets used in BLINK, adhering to the same task definitions and response formats. To assess robustness, all experiments are performed with four different random seeds. Unless noted otherwise, we set the ceiling performance  $U_j = 100$  for calculating PGF.

We first visualize the cross-task transfer matrix for each model using PGF. Each row corresponds to a source (finetuning) task, and each column denotes the target (evaluated) task (Figure 2). Figure 1 shows the PGF heatmap for Qwen-2.5-VL 32B, revealing a structured pattern of both positive and negative transfer that persists across random seeds. To highlight salient transfer relationships, we construct task transfer graphs by retaining the top 20% of strongest positive and negative PGF edges (Figure 1). To further understand the broad dynamics of task transferability, we investigate :

- ① How does transferability vary with task perception level (Low, Mid, High)?
- ② How does transferability vary with task granularity (Pixel, Crop, Image)?
- ③ How does transferability scale with model size?

#### 4.1. Task Transfer across Categories

The BLINK benchmark organizes tasks into 2 types of categories, as shown in Table 2. To study transfer dynamics at this broad level, we examine both positive and negative transferability between these categories across all model sizes. For each ordered pair of source and target categories, we compute the category-level transferability by averaging the positive and negative transfer effects between every task in the source category and every task in the target category, aggregated over four random seeds.

**Perceptual Level vs Transfer.** In the task categorization, {low-level, mid-level, high-level}, we find that finetuning on low-level tasks (Relative Depth, Relative Reflectance, Visual Correspondence) has the highest average magnitude of positive task transferability across categories, for all model sizes. In addition, low-level tasks also benefit the most on average from finetuning, achieving the highest positive malleability in all models. We present the average transferability and average malleability in Figure 3 and Figure 4 respectively.

##### Key Takeaway

Low-level tasks (Relative Depth, Relative Reflectance, Visual Correspondence) are highly positively transferable and malleable. Finetuning on low-level tasks is beneficial compared to mid and high-level tasks.

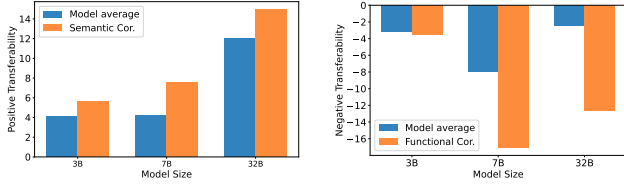
**Granularity vs Transfer.** In the task categorization, {pixel-level, crop-level, image-level}, we find that finetuning on image level tasks on average results in the highest positive transferability (Figure 3). In Figure 4, we observe that both pixel and image level tasks are malleable across model sizes.

##### Key Takeaway

Image-level tasks (Art Style, Counting, Forensic Detection, etc.) exhibit the higher positive transferability compared to pixel- and crop-level tasks. Both image-level and pixel-level tasks show higher malleability than crop-level tasks.

#### 4.2. Model Scale vs Transfer

To understand how task transferability varies with increasing model size, we analyze the average positive and negative task transferability across all tasks for each model in Figure 5. As expected, as model size increases, the average positive transferability also increases. This finding aligns with the intuition that increased representational capacity allows models to capture more generalizable features, leading to better transfer of beneficial knowledge across diverse tasks. However, there is no consistent trend with the average negative transferability. We provide detailed PGF heatmaps for all model sizes in the supplementary material.



(a) Positive task transferability across model sizes. (b) Negative task transferability across model sizes.

Figure 5. Task transferability trends across model sizes in Qwen-2.5-VL. As expected, as model size increases, the average positive transferability increases.

### Key Takeaway

The magnitude of positive transferability and malleability increases with model size.

## 5. Additional Results

In the previous sections, we examined task transfer from a broader perspective. Here, we shift to a more granular view and analyze how small clusters of mutually influential tasks emerge within the broader transfer landscape, and how distinct categories of transfer tendencies (*task personas*) characterize the roles that individual tasks play during fine-tuning.

### 5.1. Cliques of Cooperation

The improvements across tasks are not uniformly distributed and instead exhibit structured clusters of mutual influence. To formalize this observation, we define the notion of a *task clique* within the transfer graph.

**Definition 3 (Task Clique)** Let  $\{T_k\}_{k=1}^N$  denote the set of all tasks, and let  $\mu_{i \rightarrow j}$  denote a transferability score from source task  $T_i$  to target task  $T_j$ . A subset of tasks  $C \subseteq \{T_k\}_{k=1}^N$  is said to form a **task clique** if, for all ordered pairs of distinct tasks  $(T_i, T_j)$  with  $T_i, T_j \in C$  and  $i \neq j$ , the induced transfer values  $\mu_{i \rightarrow j}$  exhibit consistent sign (all positive or all negative). Tasks that mutually induce positive transfer form a *Positive Clique*, while those that mutually induce negative transfer form a *Negative Clique*.

To assess whether the extracted cliques are stable across seeds, we perform Wilcoxon tests and identify multiple statistically significant cliques of varying sizes across models. In the smaller variants (3B and 7B), the largest cliques comprise 3–4 tasks, while in Qwen-2.5-VL 32B, we observe a maximal positive clique of size 9, as illustrated in Figure 6. Detailed clique statistics and additional examples are provided in the supplementary material.

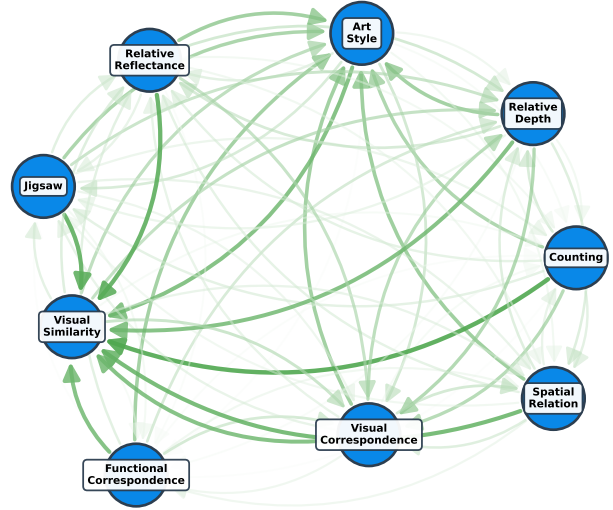


Figure 6. Positive clique of size 9 from Qwen-2.5-VL 32B.

### 5.2. Task Personas

To uncover characteristic transfer behaviors, we group tasks into distinct *personas*: source tasks that consistently help or hinder others (*Donors* and *Pirates*), and target tasks that readily absorb or resist transfer (*Sponges* and *Sieves*).

**Donors and Pirates.** Tasks which consistently exhibit a higher magnitude of positive transferability than the model average positive transferability across the model sizes are called *Donor Tasks*. Similarly, tasks that consistently induce a higher magnitude of negative transfer than the model average negative transferability across the model sizes are called *Pirate Tasks*. We identify *Semantic Correspondence* as a donor task and *Functional Correspondence* as a pirate task. Unpaired t-tests over transferability values across seeds validate that *Semantic Correspondence* is a statistically significant donor task across all models ( $p < 0.01$  across models), whereas *Functional Correspondence* is a significant pirate task in both 3B and 7B variants ( $p < 0.05$ ).

**Sponges and Sieves.** Tasks that consistently exhibit above average positive malleability across model sizes are classified as *Sponge Tasks*. On the other hand, tasks that consistently exhibit a higher magnitude of negative malleability than the model average across model sizes are classified as *Sieve Tasks*. We identify multiple *Sponge* tasks: *Visual Similarity*, *Relative Depth* and *Relative Reflectance*, whereas *Forensic Detection* emerges as the sole *Sieve* task. We conduct unpaired t-tests across seeds and find *Visual Similarity* and *Relative Depth* to be statistically significant *Sponge* tasks across models ( $p < 0.001$ ). *Forensic Detection* is a significant *Sieve* in both 3B and 32B variants ( $p < 0.005$ ).

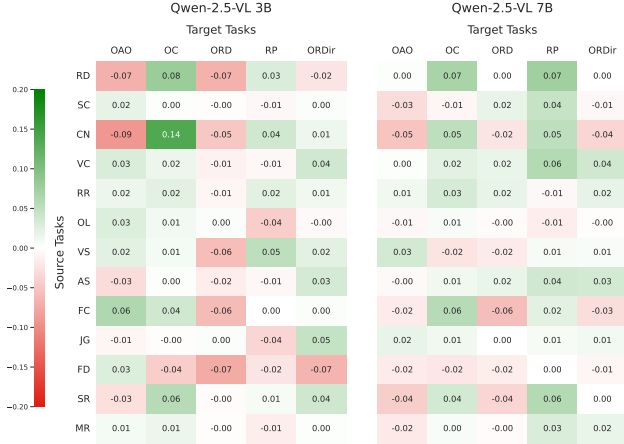


Figure 7. PGF heatmaps for Qwen-2.5-VL 3B (left) and 7B (right) models across the VSI benchmark. Consistent with previous findings, *Relative Reflectance* and *Forensic Detection* emerge as donor task and pirate task, respectively.

### 5.3. Transfer to Spatio-Temporal Tasks

To study the effects of perception task transfer to video-based tasks, we evaluate the finetuned checkpoints on VSI Bench [26] which contains a series of spatio-temporal tasks<sup>2</sup>, such as Object Counting, Object Appearance Order and Route Planning. The benchmark comprises over 5,000 question-answer pairs derived from nearly 288 egocentric indoor videos drawn from public 3D-scene datasets. It evaluates how well multimodal models perceive, recall, and reason about spatial layouts from egocentric video. We study this cross-modal transfer in the Qwen-2.5-VL 3B and 7B variants and limit the analysis to the following tasks: Object Appearance Order (OAO), Object Counting (OC), Object Relative Distance (ORD), Route planning (RP), Object Relative Direction (ORDir). The results are shown in Figure 7. Consistent with our previous findings, we note that *Relative Reflectance* emerges as a donor task and *Forensic Detection* acts as a pirate task. Moreover, we identify *Object Counting* as a sponge task while *Object Appearance Order* and *Object Relative Distance* act as sieve tasks.

#### Key Takeaway

Image-level perception tasks induce positive transfer to video-based tasks as well, demonstrating consistent trends in task transfer.

### 5.4. Data Selection with PGF

Lastly, we demonstrate how Perfection Gap Factor can guide dataset selection in the absence of task-specific training data. We consider a setting where we aim to optimize performance on some task  $T$  for which no training data is available. In-

<sup>2</sup>We do not include size estimation tasks from VSI-Bench in our analysis.

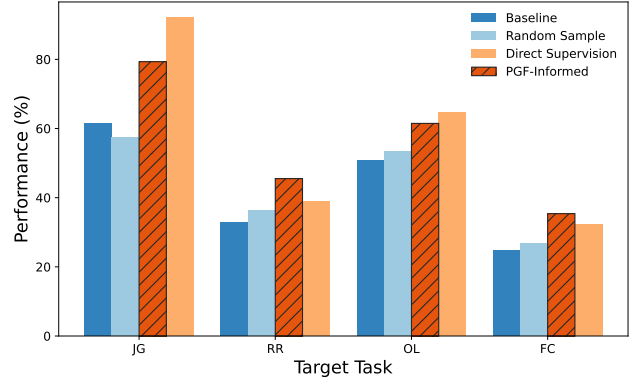


Figure 8. Performance comparison under different dataset selection strategies. PGF-informed mixtures consistently outperform random mixtures and even surpass direct supervision in two cases.

stead, we have access to datasets from several related tasks and seek an optimal mixture to improve performance on  $T$ . We propose using the most transferable tasks to  $T$  (above a certain threshold) using the PGF metric. We compare the PGF-informed dataset mixtures against randomly sampled mixture trained to optimize  $T$ . The baseline model and  $\mathcal{M}(T)$  are considered the lower and upper bound respectively. The results are presented in Figure 8. Although we limit our experiments to Qwen-2.5-VL 7B, we consistently find that PGF-informed data selection leads to better performance across multiple target tasks, demonstrating its effectiveness in guiding data selection. In case of Jigsaw and Object Localisation, PGF-informed data selection even outperforms finetuning directly on the target task itself. We note that this experiment is a preliminary finding and is included only to illustrate the potential of PGF for dataset selection. A comprehensive study of PGF-guided dataset mixtures is out of scope for this work and will be pursued in future research.

#### Key Takeaway

When lacking supervised data, PGF-informed data selection can inform alternative dataset designs which can match and even exceed the performance compared to direct finetuning.

## 6. Discussion

Our analysis of task transferability in Vision–Language Models (VLMs) reveals a rich structure in how perception capabilities interact under finetuning. Below, we unpack the broader implications of these findings, acknowledge key limitations, and propose promising directions for future work.

**Implications.** The emergence of a structured task transfer graph, characterized by cliques, personas, and scale-dependent patterns, suggests that VLMs do not treat percep-

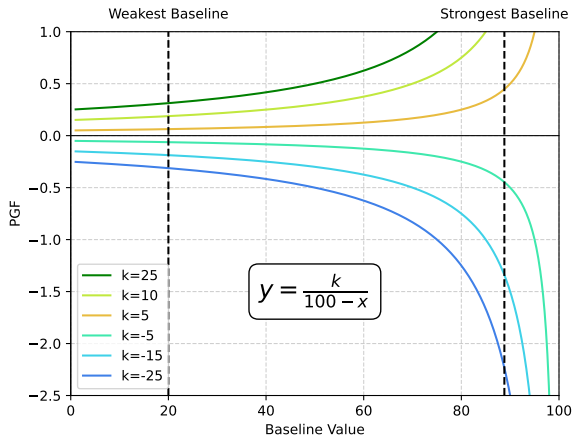


Figure 9. Behavior of PGF as a function of baseline accuracy ( $x$ ) and change after finetuning ( $k$ ).

tion tasks as independent learnings, but rather internalize them through shared or competing representational substructures. For instance, the consistent identification of low-level tasks (e.g., Relative Depth, Relative Reflectance) as strong sponges implies that early-stage visual features are highly reusable and adaptable across a wide range of downstream perception tasks. This supports the hypothesis that VLMs can benefit from hierarchical visual processing pathways. From a practical standpoint, these insights directly inform finetuning dataset design. We provide an early example of such a practical use-case. Notably, the fact that PGF-guided data selection can surpass direct target-task finetuning underscores the practical utility of our framework. Lastly, we note that our framework allows the discovery of negative cliques. This provides a unique understanding on the nature of deteriorative relationships between the considered tasks. An example of a negative clique is presented in Figure 10.

**Behavior of Perfection Gap Factor (PGF).** We further analyze the behavior of PGF. Figure 9 illustrates how 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  $m = 200$  qns,  $\text{PGF}_{\min} = -199$ . The worst-case deterioration therefore grows linearly with  $m$ .

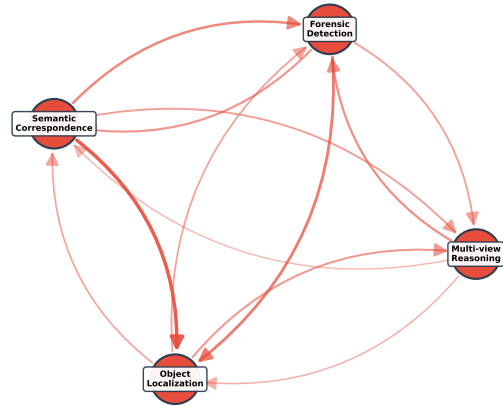


Figure 10. A negative clique of size 4 from Qwen-2.5-VL 32B.

- **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.

**Limitations.** The observations in this work come from comprehensive empirical analysis; however, it has certain limitations which can be interesting directions of future work. Our analysis is based mainly on benchmarks that model tasks in terms of multiple choice questions. This format can restrict the output space and suppress failure modes (or transfer patterns) that emerge in open-ended generation. Exploring open-ended generation for visual tasks would be a promising future direction. Besides, extending the studies to newer models will help understand the generalizability as well as evolution of architectures as pertains to their capabilities.

## 7. Conclusion

In this work, we present the first systematic analysis of perception task transfer in vision-language models. To facilitate this analysis, we introduce a new metric called Perfection Gap Factor, which helps us quantify perception task transfer in VLMs. Through experiments with three state-of-the-art VLMs, we study how finetuning on a source task impacts zero-shot performance on other tasks. Our analysis reveals several key insights. Firstly, we note that positive task transferability increases with model size. Secondly, we identify distinct cliques of mutually beneficial and mutually detrimental tasks. Lastly, we investigate inter-task interactions and characterize them as task personas. This analysis provides actionable insights into how task interactions shape model behavior, guiding the development of finetuning strategies to enhance general-purpose VLMs.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning, 2022. 3
- [3] Tianwei Chen, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Hajime Nagahara. Learning more may not be better: Knowledge transferability in vision-and-language tasks. *Journal of Imaging*, 10(12):300, 2024. 2, 3
- [4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023. 3
- [5] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 1, 3, 4
- [6] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 2
- [7] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024. 3
- [8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 3, 5
- [9] Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seung-gone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*, 2025. 3
- [10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 1
- [11] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 1
- [12] 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. 2
- [13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 1
- [14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1
- [15] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2
- [16] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. 2
- [17] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021. 1, 2
- [18] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021. 1, 2
- [19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. Pmlr, 2021. 1
- [20] M Moein Shariatnia, Rahim Entezari, Mitchell Wortsman, Olga Saukh, and Ludwig Schmidt. How well do contrastively trained models transfer? In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. 2, 3
- [21] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. 1
- [22] Anthony Meng Huat Tiong, Junqi Zhao, Boyang Li, Junnan Li, Steven CH Hoi, and Caiming Xiong. What are we measuring when we evaluate large vision-language models? an analysis of latent factors and biases. *arXiv preprint arXiv:2404.02415*, 2024. 2, 3
- [23] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1
- [24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024. 1, 4
- [25] Anne Wu, Kianté Brantley, and Yoav Artzi. A surprising failure? multimodal llms and the nlvr challenge, 2024. 3
- [26] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 7
- [27] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1, 2

- [28] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. [2](#), [3](#)
- [29] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#)