Résumé: Vision-Language Models (VLMs) have revolutionized open-vocabulary recognition, yet adapting them to specialized domains under data-scarce conditions remains a significant challenge. This thesis investigates efficient strategies to bridge the gap between inductive fine-tuning and transductive inference, focusing on minimizing supervision and computational costs. We first address inductive few-shot learning with CLIP-LoRA, a parameter-efficient fine-tuning method. By injecting trainable low-rank matrices into the encoder, it outperforms existing approaches while eliminating dataset-specific hyperparameter tuning. For scenarios restricted to single-sample inference, we propose MTA (MeanShift for Test-time Augmentation). Formulated as a robust variation of the MeanShift algorithm, this training-free method refines predictions through iterative mode-seeking and automatic importance weighting of augmented views. Expanding into the transductive paradigm, we introduce TransCLIP. This framework explicitly models the underlying structure of unlabeled test batches by fitting a Gaussian Mixture Model (GMM), constrained by a text-driven Kullback-Leibler (KL) regularization term. This dual objective dynamically aligns visual statistics with textual priors, achieving substantial gains in both zero-shot and few-shot settings. Finally, to address realistic deployment challenges, we propose StatA (Statistical Anchor). We demonstrate that existing transductive methods fail under non-i.i.d. data streams and introduce a regularization technique based on the KL divergence with respect to initial text priors. Collectively, this thesis provides a suite of transparent, low-compute algorithms that, through the use of fixed hyperparameters, ensure operational simplicity and robust performance across challenging and diverse visual recognition benchmarks.