



Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences



Fraunhofer
IAIS



Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

R&D Project Defense

March 13, 2025

Shinas Shaji

Advisors

Prof. Dr. Sebastian Houben (Hochschule Bonn-Rhein-Sieg, Fraunhofer IAIS)

Santosh Thoduka M.Sc. (Fraunhofer IAIS)

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation
Models to Learn Lightweight Models for Robotic Vision Tasks



Evaluation of Few-Shot Transfer of
Vision-Language Foundation Models to Learn
Lightweight Models for Robotic Vision Tasks

R&D Project Defense

March 13, 2025

Shinas Shaji

Advisors

Prof. Dr. Sebastian Houben (Hochschule Bonn-Rhein-Sieg, Fraunhofer IAIS)
Santosh Thoduka M.Sc. (Fraunhofer IAIS)

1. Introduction

1.1 Background

1.2 Problem Statement

2. Related Work

2.1 Background

2.2 Research Gaps

3. Proposed Approach

4. Methodology

4.1 Datasets

4.2 Models Chosen

4.3 Experimental Design

5. Evaluation

5.1 CIFAR-10 Experiments

5.2 Seven-Point Checklist Dermatology Experiments

6. Conclusions

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Introduction

| |
|---|
| 1. Introduction |
| 1.1 Background |
| 1.2 Problem Statement |
| 2. Related Work |
| 2.1 Background |
| 2.2 Research Gaps |
| 3. Proposed Approach |
| 4. Methodology |
| 4.1 Datasets |
| 4.2 Models Chosen |
| 4.3 Experimental Design |
| 5. Evaluation |
| 5.1 CIFAR-10 Experiments |
| 5.2 Seven-Point Checklist Dermatology Experiments |
| 6. Conclusions |

Introduction

Vision-Language Models (VLMs)

- Neural networks that process both **images** and **text**
- Like Large Language Models (LLMs):
 - Learn **general visual-textual understanding** from pre-training ¹
 - Then aligned to human preferences and **instruction-following**
- Shown to be able to **adapt** to new tasks without extensive task-specific training ², hence quite **generalizable**

¹ A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila & T. Zhang, Eds., ser. *Proceedings of Machine Learning Research*, vol. 139, PMLR, Jul. 2021, pp. 8748–8763

² P. Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Introduction
 - └ Background
 - └ Introduction

- I know that most of you are already familiar with multi-modal models, being in a room of researchers in multimodal foundation models, but perhaps there are people in the audience who are not familiar with the field. So we start with a brief introduction to some concepts we will be needing today
- VLMs combine computer vision and natural language processing capabilities in a single neural network
- Pre-training occurs on web-scale image-text datasets (e.g., CLIP was trained on 400M image-text pairs). In this way, models learn to understand the relationship between images and text and also predict one from the other
- Unlike traditional CV models trained for specific tasks, VLMs learn broader understanding that transfers to many tasks
- Similar to how ChatGPT can answer questions on many topics, VLMs can handle multiple visual tasks

- Neural networks that process both **images** and **text**
- Like Large Language Models (LLMs):
 - Learn **general visual-textual understanding** from pre-training ¹
 - Then aligned to human preferences and **instruction-following**
- Shown to be able to **adapt** to new tasks without extensive task-specific training ², hence quite **generalizable**

¹ A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila & T. Zhang, Eds., ser. *Proceedings of Machine Learning Research*, vol. 139, PMLR, Jul. 2021, pp. 8748–8763

² P. Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023

Introduction

Few-Shot Transfer

- Teaching a **generalizable model** new tasks by showing it a few **examples**
- Model 'learns' to recognize patterns from these examples
- Can then apply this 'learning' to new, unseen instances

Prompt: A [DOG] has droopy ears and is often fluffy. This is a [DOG]:



Figure 1: This is a [DOG].
Image from Wikipedia, Link



Figure 2: Is this a [DOG]?
Image from Wikipedia, Link

Prompt: Is this a [DOG]?
Expected Answer: Yes

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Introduction
- └ Background
- └ Introduction

- Few-shot transfer is different from traditional machine learning that requires many examples
- The model doesn't actually learn in the traditional sense; it leverages knowledge from pre-training
- In this example, we show the model one example of a dog (a basset hound) and ask it to identify another dog
- This works because the model already knows what dogs look like from pre-training
- The quotation marks around 'learns' emphasize that it's not learning from scratch. I like to think of it as somehow narrowing the tree of possible responses from a point

Introduction
Few-Shot Transfer

- Teaching a **generalizable model** new tasks by showing it a few **examples**
- Model 'learns' to recognize patterns from these examples
- Can then apply this 'learning' to new, unseen instances

Prompt: A [DOG] has droopy ears and is often fluffy. This is a [DOG]:




Figure 1: This is a [DOG].
Image from Wikipedia, Link




Figure 2: Is this a [DOG]?
Image from Wikipedia, Link

Prompt: Is this a [DOG]?
Expected Answer: Yes

Motivation

Few-Shot Transfer

Fine-tuning

- Requires significant computational resources, modifies model parameters
- Needs **large(r) amounts** of **labeled data**
- Can lead to catastrophic forgetting
- Refers to fine-tuning a pre-trained / instruction-tuned model on a specific task

Few-shot Transfer

- Uses **‘few’ examples** or natural language **descriptions**
- No model parameters are updated
- Potentially more practical for real-world applications ^a
- Can be less effective for complex tasks

^aT. Brown et al., Language Models are Few-Shot Learners, in *Advances in Neural Information Processing Systems*, H. Larochelle et al., Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Introduction
 - └ Background
 - └ Motivation

- Fine-tuning is the traditional approach for adapting pre-trained models to new tasks
- For VLMs, fine-tuning can require multiple high-end GPUs and days of compute time
- Catastrophic forgetting means the model may lose previously learned capabilities
- Few-shot transfer is much more efficient - we can use the model "as is"
- An important distinction: fine-tuning modifies the model weights, few-shot transfer doesn't
- The trade-off is that few-shot performance may not match fine-tuned performance for complex tasks

Problem Statement

Dataset Labeling for Computer Vision Tasks

Challenge: Creating labeled datasets to train specialized models for computer vision tasks is **time-consuming** and **expensive**³, but VLMs are generalizable

Constraint: However, VLMs are too **computationally intensive** for direct deployment on resource-constrained environments (e.g., robots)

Opportunity: VLMs could potentially automate label generation (**pseudolabels**) to train **downstream** models

Research Question: Can VLMs be transferred to generate **pseudolabels** for computer vision tasks to train **lightweight** downstream models?

³ J. Deng et al., ImageNet: A Large-Scale Hierarchical Image Database, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Introduction
 - └ Problem Statement
 - └ Problem Statement

- The annotation of the ImageNet dataset required crowdsourcing to be feasible
- Specialized datasets like medical imaging can be even more expensive and require expert annotators
- VLMs require significant computational resources - multiple high-end GPUs with significant amounts of memory
- Most robots have limited computational capabilities - often a single low-power GPU or CPU
- Pseudolabels are automatically generated annotations that can be used in place of human-created labels
- The key insight: we don't need to deploy the large model - we use it to train a smaller, more efficient model

Problem Statement

Dataset Labeling for Computer Vision Tasks

Challenge: Creating labeled datasets to train specialized models for computer vision tasks is **time-consuming** and **expensive**³, but VLMs are generalizable

Constraint: However, VLMs are too **computationally intensive** for direct deployment on resource-constrained environments (e.g., robots)

Opportunity: VLMs could potentially automate label generation (**pseudolabels**) to train **downstream** models

Research Question: Can VLMs be transferred to generate **pseudolabels** for computer vision tasks to train **lightweight** downstream models?

³ J. Deng et al., ImageNet: A Large-Scale Hierarchical Image Database, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255

1. Introduction

1.1 Background

1.2 Problem Statement

2. Related Work

2.1 Background

2.2 Research Gaps

3. Proposed Approach

4. Methodology

4.1 Datasets

4.2 Models Chosen

4.3 Experimental Design

5. Evaluation

5.1 CIFAR-10 Experiments

5.2 Seven-Point Checklist Dermatology Experiments

6. Conclusions

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Related Work

| |
|---|
| 1. Introduction |
| 1.1 Background |
| 1.2 Problem Statement |
| 2. Related Work |
| 2.1 Background |
| 2.2 Research Gaps |
| 3. Proposed Approach |
| 4. Methodology |
| 4.1 Datasets |
| 4.2 Models Chosen |
| 4.3 Experimental Design |
| 5. Evaluation |
| 5.1 CIFAR-10 Experiments |
| 5.2 Seven-Point Checklist Dermatology Experiments |
| 6. Conclusions |

Related Work

Vision-Language Models

Vision-Language Model Classes

- **Alignment models:** Generate unified text-image embeddings (CLIP ^a, FLAVA)
- **Generative models:** Generate text conditioned on multimodal inputs (Flamingo, Frozen ^b, MiniCPM ^c, GPT-4o, Claude, etc.)

^aA. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila & T. Zhang, Eds., ser. *Proceedings of Machine Learning Research*, vol. 139, PMLR, Jul. 2021, pp. 8748–8763

^bM. Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models, in *Advances in Neural Information Processing Systems*, M. Ranzato et al., Eds., vol. 34, Curran Associates, Inc., 2021, pp. 200–212

^cY. Yao et al., MiniCPM-V: A GPT-4V Level MLLM on Your Phone, *CoRR*, vol. abs/2408.01800, 2024. arXiv: 2408.01800 [cs.CV]

Prompting Techniques: Crafting prompts to improve task performance ^a

- In-context learning: Providing **examples** in context ^b
- Chain-of-thought prompting for complex **reasoning** ^c

^aP. Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023

^bT. Brown et al., Language Models are Few-Shot Learners, in *Advances in Neural Information Processing Systems*, H. Larochelle et al., Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901

^cJ. Wei et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in *Advances in Neural Information Processing Systems*, S. Koyejo et al., Eds., vol. 35, Curran Associates, Inc., 2022, pp. 24 824–24 837

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Related Work

└ Background

└ Related Work

- Vision-Language Models (VLMs) have progressed significantly since the introduction of CLIP
- Alignment models, such as CLIP, are designed to learn joint representations but do not generate textual outputs. What are joint representations? Text and images are processed into vectors or embeddings such that for similar concepts, the embeddings are close to each other in vector space
- In contrast, generative models are capable of producing text responses based on visual inputs
- Contemporary VLMs often incorporate separate pre-trained components for vision and language, connected through adapters or connectors
- Additionally, unified architectures that process both modalities simultaneously have emerged, enhancing the integration of vision and language tasks
- A pivotal advancement facilitating few-shot transfer is the conceptualization of vision tasks as text generation tasks. This paradigm allows us to articulate tasks using natural language, eliminating the need for specialized architectures
- Prompting techniques are crucial for adapting models to new tasks without retraining
- In-context learning allows models to learn from examples without parameter updates

Related Work
Vision-Language Models
Vision-Language Model Classes

- **Alignment models:** Generate unified text-image embeddings (CLIP ^a, FLAVA)
- **Generative models:** Generate text conditioned on multimodal inputs (Flamingo, Frozen ^b, MiniCPM ^c, GPT-4o, Claude, etc.)

Prompting Techniques: Crafting prompts to improve task performance ^a

- In-context learning: Providing **examples** in context ^b
- Chain-of-thought prompting for complex **reasoning** ^c

^aP. Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023

^bT. Brown et al., Language Models are Few-Shot Learners, in *Advances in Neural Information Processing Systems*, H. Larochelle et al., Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901

^cJ. Wei et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in *Advances in Neural Information Processing Systems*, S. Koyejo et al., Eds., vol. 35, Curran Associates, Inc., 2022, pp. 24 824–24 837

Related Work

Applications and Datasets

Applicability of VLMs

- Large-scale pre-training enables **generalization** to various tasks
- Traditional DNNs trained for **specific** tasks

LLMs for Data Annotation

- LLMs used to generate multimodal instruction-following data ^a
- Seen to outperform crowd-workers in **data annotation** tasks ^b

^aH. Liu et al., Visual instruction tuning, in *Advances in Neural Information Processing Systems*, A. Oh et al., Eds., vol. 36, Curran Associates, Inc., 2023, pp. 34 892–34 916

^bF. Gilaridi et al., ChatGPT outperforms crowd workers for text-annotation tasks, *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, e2305016120, 2023

Key Datasets: Various datasets exist for various vision tasks.

- ImageNet, CIFAR-10 ^a, GTSDb ^b: Object recognition, detection
- Microsoft COCO: Object detection, segmentation, captioning
- Derm7Pt ^c: Specialized dermatology dataset
- MVTec: Anomaly detection

^aA. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, M.S. thesis, Department of Computer Science, University of Toronto, 2009

^bS. Houben et al., Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark, in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8

^cJ. Kawahara et al., Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets, *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Related Work

└ Background

└ Related Work

- VLMs are used in a wide range of vision tasks, leveraging their ability to generalize from large-scale pre-training
- Traditional DNNs are typically trained on specific datasets for specific tasks, limiting their generalization
- LLMs have been used to generate multimodal instruction-following data to train VLMs, enhancing dataset creation
- LLMs have been seen to outperform human annotators, providing more consistent and scalable data annotation
- Few-shot transfer capabilities of VLMs are less studied compared to LLMs in NLP
- We acknowledge important VLM limitations:
- Difficulty with spatial relationships and perspective
- Inherited issues from LLMs like hallucinations
- Biases from pre-training data making model ignore visual context
- Something to note: this is a large field of research, and most of the research was done in the last 4-5 years

Related Work

Applications and Datasets

Applicability of VLMs

- Large-scale pre-training enables **generalization** to various tasks
- Traditional DNNs trained for **specific** tasks

LLMs for Data Annotation

- LLMs used to generate multimodal instruction-following data ^a
- Seen to outperform crowd-workers in **data annotation** tasks ^b

^aH. Liu et al., Visual instruction tuning, in *Advances in Neural Information Processing Systems*, A. Oh et al., Eds., vol. 36, Curran Associates, Inc., 2023, pp. 34 892–34 916

^bF. Gilaridi et al., ChatGPT outperforms crowd workers for text-annotation tasks, *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, e2305016120, 2023

Key Datasets: Various datasets exist for various vision tasks.

- ImageNet, CIFAR-10 ^a, GTSDb ^b: Object recognition, detection
- Microsoft COCO: Object detection, segmentation, captioning
- Derm7Pt ^c: Specialized dermatology dataset
- MVTec: Anomaly detection

^aA. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, M.S. thesis, Department of Computer Science, University of Toronto, 2009

^bS. Houben et al., Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark, in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8

^cJ. Kawahara et al., Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets, *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019

Related Work

Research Gaps

Gaps Addressed in Our Work

- **Few-shot** transfer of **VLMs** (with in-context learning) **less explored** than NLP counterparts ⁴
- **Dataset annotation** applications ⁵:
 - Use of VLMs for dataset annotation **less explored**
 - Lack of analyses on **downstream vision models** from VLM-generated pseudolabels

⁴ P. Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023

⁵ Z. Tan et al., Large Language Models for Data Annotation and Synthesis: A Survey, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan et al., Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 930–957. arXiv: 2402.13446 [cs.CL]



2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Related Work
 - └ Research Gaps
 - └ Related Work

Related Work

Research Gaps

Gaps Addressed in Our Work

- Few-shot transfer of VLMs (with in-context learning) **less explored** than NLP counterparts ⁴
- Dataset annotation applications ⁵:
 - Use of VLMs for dataset annotation **less explored**
 - Lack of analyses on **downstream vision models** from VLM-generated pseudolabels

⁴ Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023.
⁵ Tan et al., Large Language Models for Data Annotation and Synthesis: A Survey, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan et al., Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 930–957. arXiv: 2402.13446 [cs.CL]

1. Introduction

1.1 Background

1.2 Problem Statement

2. Related Work

2.1 Background

2.2 Research Gaps

3. Proposed Approach

4. Methodology

4.1 Datasets

4.2 Models Chosen

4.3 Experimental Design

5. Evaluation

5.1 CIFAR-10 Experiments

5.2 Seven-Point Checklist Dermatology Experiments

6. Conclusions

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Proposed Approach

| | |
|---|--|
| 1. Introduction | |
| 1.1 Background | |
| 1.2 Problem Statement | |
| 2. Related Work | |
| 2.1 Background | |
| 2.2 Research Gaps | |
| 3. Proposed Approach | |
| 4. Methodology | |
| 4.1 Datasets | |
| 4.2 Models Chosen | |
| 4.3 Experimental Design | |
| 5. Evaluation | |
| 5.1 CIFAR-10 Experiments | |
| 5.2 Seven-Point Checklist Dermatology Experiments | |
| 6. Conclusions | |

Proposed Approach

Evaluating VLMs for Pseudolabel Generation

Approach: Evaluate VLMs on generating **accurate pseudolabels** for image classification datasets under various **zero-shot** and **few-shot** transfer conditions

Key Research Aspects

1. How does the **number of examples** (few-shot vs. zero-shot) affect pseudolabel quality?
2. What are the **computational requirements** for practical application?
3. How effective are the **downstream models** trained on pseudolabels?

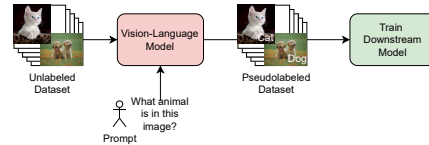


Figure 3: Using VLMs to generate pseudolabels for downstream model training

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Proposed Approach

└ Proposed Approach

- Zero-shot means only describing the task in natural language; few-shot means providing examples
- We tested multiple prompting strategies to find the most effective approach
- We examined both common visual tasks (CIFAR-10) and specialized tasks (dermatology) to see if few-shot transfer is generalizable
- This design helps us understand when VLMs can be reliably transferred for automatic labeling
- For computational requirements, we measured inference time, memory usage, and scaling properties
- Downstream models were evaluated against models trained on ground truth labels

Proposed Approach

Evaluating VLMs for Pseudolabel Generation

Approach: Evaluate VLMs on generating **accurate pseudolabels** for image classification datasets under various **zero-shot** and **few-shot** transfer conditions

Key Research Aspects

1. How does the **number of examples** (few-shot vs. zero-shot) affect pseudolabel quality?
2. What are the **computational requirements** for practical application?
3. How effective are the **downstream models** trained on pseudolabels?



1. Introduction

1.1 Background

1.2 Problem Statement

2. Related Work

2.1 Background

2.2 Research Gaps

3. Proposed Approach

4. Methodology

4.1 Datasets

4.2 Models Chosen

4.3 Experimental Design

5. Evaluation

5.1 CIFAR-10 Experiments

5.2 Seven-Point Checklist Dermatology Experiments

6. Conclusions

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Methodology

| | |
|---|--|
| 1. Introduction | |
| 1.1 Background | |
| 1.2 Problem Statement | |
| 2. Related Work | |
| 2.1 Background | |
| 2.2 Research Gaps | |
| 3. Proposed Approach | |
| 4. Methodology | |
| 4.1 Datasets | |
| 4.2 Models Chosen | |
| 4.3 Experimental Design | |
| 5. Evaluation | |
| 5.1 CIFAR-10 Experiments | |
| 5.2 Seven-Point Checklist Dermatology Experiments | |
| 6. Conclusions | |

Datasets Used

CIFAR-10 Dataset

Dataset Characteristics

- 60,000 RGB images (32x32 pixels)
- 10 **balanced** categories
- Standard split:
 - 50,000 training
 - 10,000 test (1,000/class)
- **Common** vision task



Figure 4: Example images from CIFAR-10 ^a

^aA. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, M.S. thesis, Department of Computer Science, University of Toronto, 2009

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Methodology
└ Datasets
└ Datasets Used

Datasets Used

CIFAR-10 Dataset

- Dataset Characteristics**
- 60,000 RGB images (32x32 pixels)
 - 10 **balanced** categories
 - Standard split:
 - 50,000 training
 - 10,000 test (1,000/class)
 - **Common** vision task



Figure 4: Example images from CIFAR-10 ^a

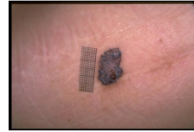
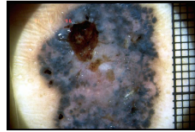
^aA. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, M.S. thesis, Department of Computer Science, University of Toronto, 2009

Datasets Used

Seven-Point Checklist Dermatology Dataset (Derm7Pt)

Dataset Characteristics

- 1,011 dermatological cases for melanoma diagnosis (as well as other skin conditions)
- **Multiple** image **modalities**:
 - Dermoscopic (standardized)
 - Clinical (varying conditions)
- Diagnoses and features are **not uniformly distributed**
- Additional task information: **describes** each **checklist feature**
- **Task-specific** dataset – likely not present in VLM pre-training data



diagnosis: basal cell carcinoma
pigment_network (PN): absent
streaks (STR): absent
pigmentation (PIG): absent
regression_structures (RS): blue areas
dots_and_globules (DaG): irregular
blue_whitish_veil (BWV): present
vascular_structures (VS): within regression
seven_point_score: 4
level_of_diagnostic_difficulty: low

Figure 5: An example case from Derm7Pt

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Methodology

Datasets

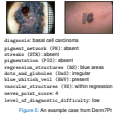
Datasets Used

- The Derm7Pt dataset is a collection of dermatological images and annotations for melanoma diagnosis, as well as other skin conditions such as seborrheic keratosis, basal cell carcinoma, types of nevi, and more
- A seven point score of greater than or equal to 3 is to be referred for specialist analysis for melanoma
- The authors provide a python library as an interface to the dataset, and perform grouping of rare cases as well as the grouping of fine-grained categorization into more coarse categories
- In the authors' paper, it is shown that incorporating clinical images as input to their CNN model improves performance

Datasets Used

Seven-Point Checklist Dermatology Dataset (Derm7Pt)

- 1,011 dermatological cases for melanoma diagnosis (as well as other skin conditions)
- **Multiple image modalities**:
 - Dermoscopic (standardized)
 - Clinical (varying conditions)
- Diagnoses and features are **not uniformly distributed**
- Additional task information: **describes each checklist feature**
- **Task-specific** dataset – likely not present in VLM pre-training data



Models Chosen

Which models did we evaluate?

- **InternVL2-8B:** 8B parameters, 8k context window
- **MiniCPM-V-2.6:** 8B parameters, **image token compression** with **perceiver-resampler** architecture
- **Pixtral 12B:** 12B parameters, 128K context window, **handles images natively**
- **Phi-3.5-Vision-Instruct:** 4B parameters, 128K context window, designed for constrained environments
- **Bio-Medical-MultiModal-Llama-3-8B-V1:** 8B parameters, **fine-tuned** multimodal adaptation of Llama-3.1-8B-Instruct for **biomedical** tasks

Notes:

- All models use towered architectures, integrating visual encoders with language models, with **integrated image** tiling, **processing** methods (except for Pixtral)
- Models chosen for claimed ability in processing **interleaved** image-text inputs

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└ Methodology

└ Models Chosen

└ Models Chosen

- InternVL2-8B and MiniCPM-V-2.6 models integrate visual and language processing components
- Phi-3.5-vision-instruct and Pixtral 12B models are designed for long-context multimodal processing
- Bio-Medical-MultiModal-Llama-3-8B-V1 is specialized for biomedical tasks
- Models selected for their ability to handle multi-image, multi-turn multimodal conversations, or in other words, interleaved text and image inputs
- Notably, all models use their own integrated image tiling and processing methods, except for the Pixtral 12B model, which processes images natively at their full resolution without tiling

Models Chosen

Which models did we evaluate?

- **InternVL2-8B:** 8B parameters, 8k context window
- **MiniCPM-V-2.6:** 8B parameters, **image token compression** with **perceiver-resampler** architecture
- **Pixtral 12B:** 12B parameters, 128K context window, **handles images natively**
- **Phi-3.5-Vision-Instruct:** 4B parameters, 128K context window, designed for constrained environments
- **Bio-Medical-MultiModal-Llama-3-8B-V1:** 8B parameters, **fine-tuned** multimodal adaptation of Llama-3.1-8B-Instruct for **biomedical** tasks

Notes:

- All models use towered architectures, integrating visual encoders with language models, with **integrated image** tiling, **processing** methods (except for Pixtral)
- Models chosen for claimed ability in processing **interleaved** image-text inputs

Experimental Design

How did we evaluate VLMs?

- **Zero-Shot Transfer:**

- Evaluated ability to perform tasks **without prior examples**
- Used basic and enhanced prompts with task-specific information and chain-of-thought reasoning

- **Few-Shot Transfer:**

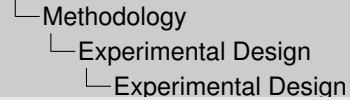
- Assessed effects of **providing examples** on model performance, resource usage
- Explored pure few-shot experiments and combinations with prompting

- **Downstream Model Training:**

- Evaluated **knowledge transfer** by training downstream models from pseudolabels
- Analyzed effects of transfer methods on downstream model performance

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks



- Zero-shot learning assessed through basic and enhanced prompts
- Few-shot learning explored with varying examples and combinations
- Downstream training focused on pseudolabel effectiveness
- Comprehensive evaluation framework used to assess VLM capabilities

- **Zero-Shot Transfer:**
 - Evaluated ability to perform tasks **without prior examples**
 - Used basic and enhanced prompts with task-specific information and chain-of-thought reasoning
- **Few-Shot Transfer:**
 - Assessed effects of **providing examples** on model performance, resource usage
 - Explored pure few-shot experiments and combinations with prompting
- **Downstream Model Training:**
 - Evaluated **knowledge transfer** by training downstream models from pseudolabels
 - Analyzed effects of transfer methods on downstream model performance

Prompting and VLM Inputs

How did we prompt the VLM?

A **Prompting Framework** was developed, that:

- Employed a consistent **chat-based** interaction format for fair comparison
- Isolated effects of different prompting components (zero-shot, few-shot, reasoning) with a **modular** prompt structure
- **Minimized biases** through systematic, deterministic and reproducible **randomization** of class lists and example ordering

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Methodology
 - └ Experimental Design
 - └ Prompting and VLM Inputs

- Prompting strategy focused on fair comparison and isolating effects of different components
- Chat-based format maintained consistency across experiments
- Standardized preprocessing and prompt templates ensured reproducibility
- Deterministic random seeds used for example selection and class list shuffling
- Core structure included base prompt, example integration, and test instance presentation

A **Prompting Framework** was developed, that:

- Employed a consistent **chat-based** interaction format for fair comparison
- Isolated effects of different prompting components (zero-shot, few-shot, reasoning) with a **modular** prompt structure
- **Minimized biases** through systematic, deterministic and reproducible **randomization** of class lists and example ordering

Prompting and VLM Inputs

How did we prompt the VLM?

Core Structure of the Prompt:

- **Base Prompt Structure:** **Guides models** with task definition and output format
- **Example Integration:** Alternating user-assistant message pairs for **few-shot examples**
- **Test Instance Presentation:** Test image presented for classification in **final turn**

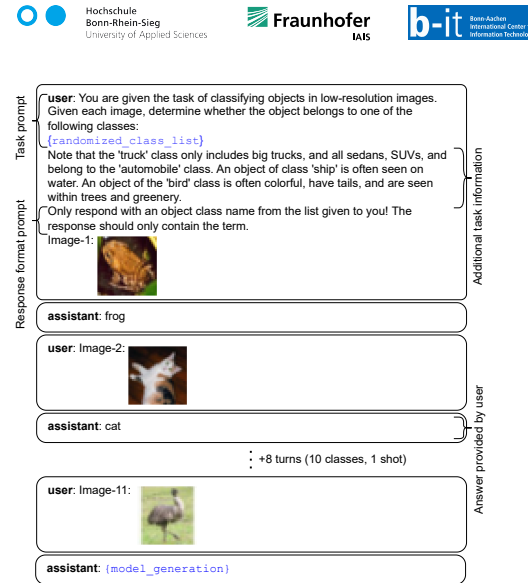


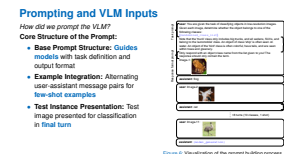
Figure 6: Visualization of the prompt building process

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- Methodology
 - Experimental Design
 - Prompting and VLM Inputs

- The example shown is for the CIFAR-10 dataset. A similar process is used for the Derm7Pt dataset, but with category-specific prompts for each category and the optional addition of clinical images
- Base prompt structure guides models with task definition, class list, reasoning requirements, and output format - Alternating user-assistant message pairs for few-shot examples – answers are provided by user - Final turn presents test image for classification – answer is provided by model



1. Introduction

1.1 Background

1.2 Problem Statement

2. Related Work

2.1 Background

2.2 Research Gaps

3. Proposed Approach

4. Methodology

4.1 Datasets

4.2 Models Chosen

4.3 Experimental Design

5. Evaluation

5.1 CIFAR-10 Experiments

5.2 Seven-Point Checklist Dermatology Experiments

6. Conclusions

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Evaluation

- 1. Introduction
 - 1.1 Background
 - 1.2 Problem Statement
- 2. Related Work
 - 2.1 Background
 - 2.2 Research Gaps
- 3. Proposed Approach
- 4. Methodology
 - 4.1 Datasets
 - 4.2 Models Chosen
 - 4.3 Experimental Design
- 5. Evaluation
 - 5.1 CIFAR-10 Experiments
 - 5.2 Seven-Point Checklist Dermatology Experiments
- 6. Conclusions

CIFAR-10 Experiments

A General Image Classification Task

Dataset Characteristics

- 60,000 RGB images (32x32 pixels)
- 10 **balanced** categories
- Standard split:
 - 50,000 training
 - 10,000 test (1,000/class)
- **Common** vision task



Figure 7: Example images from CIFAR-10 ^a

^aA. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, M.S. thesis, Department of Computer Science, University of Toronto, 2009

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Evaluation
 - └ CIFAR-10 Experiments
 - └ CIFAR-10 Experiments

- Only a short refresher

CIFAR-10 Experiments
A General Image Classification Task

Dataset Characteristics

- 60,000 RGB images (32x32 pixels)
- 10 **balanced** categories
- Standard split:
 - 50,000 training
 - 10,000 test (1,000/class)
- **Common** vision task

Figure 7: Example images from CIFAR-10 ^a



CIFAR-10 Experiments

Performance Characteristics and Insights, CIFAR-10

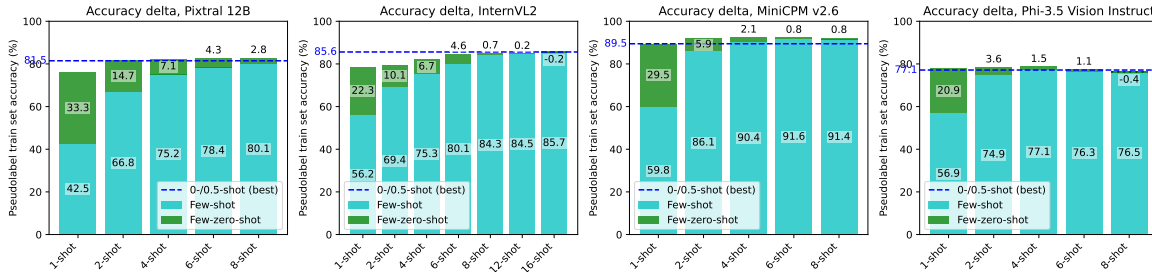


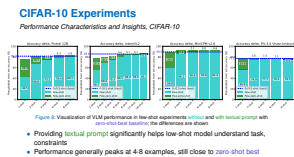
Figure 8: Visualization of VLM performance in few-shot experiments **without** and **with textual prompt** with **zero-shot-best baseline**; the differences are shown

- Providing **textual prompt** significantly helps low-shot model understand task, constraints
- Performance generally peaks at 4-8 examples, still close to **zero-shot best**

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Evaluation
 - └ CIFAR-10 Experiments
 - └ CIFAR-10 Experiments



- Without prompting, model generates predictions not in class list, leading to a large number of invalid predictions
- With enough examples, model 'learns' the class list, supporting the idea that the model 'recognizes' the task and its constraints rather than 'learn' it

Downstream Model Analysis

Performance Characteristics and Insights, CIFAR-10

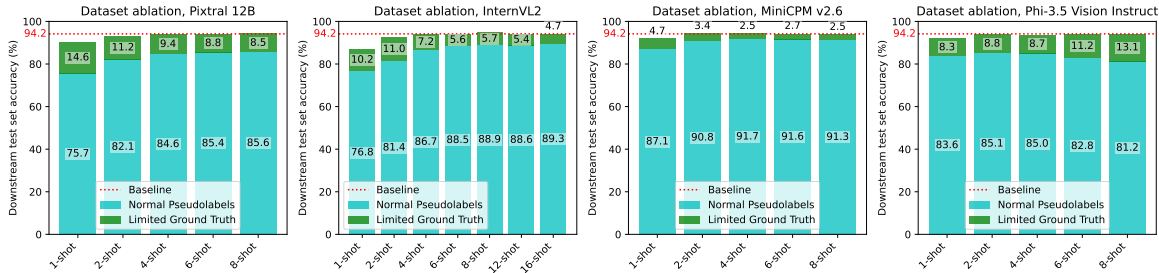


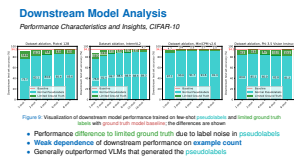
Figure 9: Visualization of downstream model performance trained on few-shot pseudolabels and limited ground truth labels with ground truth model baseline; the differences are shown

- Performance difference to limited ground truth due to label noise in pseudolabels
- Weak dependence of downstream performance on example count
- Generally outperformed VLMs that generated the pseudolabels

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

- └ Evaluation
 - └ CIFAR-10 Experiments
 - └ Downstream Model Analysis



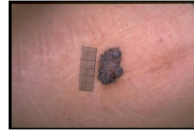
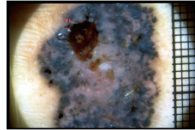
- Downstream models consistently outperformed VLMs due to ImageNet pre-training and focus on pattern recognition
- Label quality proved more important than dataset size in training. In few-shot experiments, 33% of the labels were invalid and hence is equivalent to a reduction in dataset size by 33%
- However, the performance was still acceptable, with only a minor drop in performance compared to a model trained on ground truth labels with the same invalid labels taken out. Practically utility – can use imperfect data
- Downstream model performance has weak dependence on example count, meaning both dataset size and number of shots - Downstream models consistently outperformed VLMs due to ImageNet pre-training and focus on pattern recognition

Dermatology Experiments

Experiment Setup and Evaluation Framework

Dataset Characteristics

- 1,011 dermatological cases for melanoma diagnosis (as well as other skin conditions)
- **Multiple** image **modalities**:
 - Dermoscopic (standardized)
 - Clinical (varying conditions)
- Diagnoses and features are **not uniformly distributed**
- Additional task information: **describes** each **checklist feature**
- **Task-specific** dataset – likely not present in VLM pre-training data



diagnosis: basal cell carcinoma
pigment_network (PN): absent
streaks (STR): absent
pigmentation (PIG): absent
regression_structures (RS): blue areas
dots_and_globules (DaG): irregular
blue_whitish_veil (BWV): present
vascular_structures (VS): within regression
seven_point_score: 4
level_of_diagnostic_difficulty: low

Figure 10: An example case from Derm7Pt

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Evaluation

Seven-Point Checklist Dermatology Experiments

Dermatology Experiments

- Only a short refresher
- We have two tasks: one that tries to predict the diagnosis directly from the images, and another that predicts each of the seven checklist features from the images
- Idea: Maybe the model finds it difficult to predict the condition directly, so maybe it is better to predict the features first?

Dermatology Experiments

Performance in Direct Condition Diagnosis

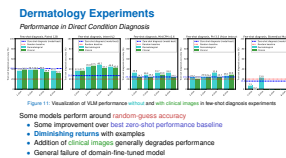
2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Evaluation

Seven-Point Checklist Dermatology Experiments

Dermatology Experiments



- Direct diagnosis showed some promise with few-shot learning
- Few-shot performance generally peaks at 2-4 examples
- Only best zero-shot performance results shown here, but reasoning requirements generally decreased model performance
- Only Pixtral 12B had zero-shot performance notable better than random guessing, but the other models caught up with few-shot prompting
- The domain-specific Biomedical Multimodal Model struggled, and this is likely because the fine-tuning caused regressions in the instruction-following capabilities of the model. This is somewhat of a known issue, and goes along the lines of catastrophic forgetting – but this is something that a few-shot or zero-shot transferred model may not suffer from

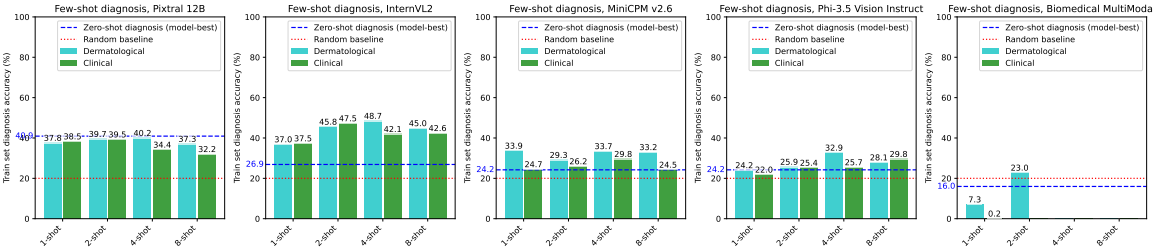


Figure 11: Visualization of VLM performance *without* and *with* clinical images in few-shot diagnosis experiments

Some models perform around **random-guess accuracy**

- Some improvement over **best zero-shot performance baseline**
- **Diminishing returns** with examples
- Addition of **clinical images** generally degrades performance
- General failure of domain-fine-tuned model

Dermatology Experiments

Performance in Structured Feature Understanding

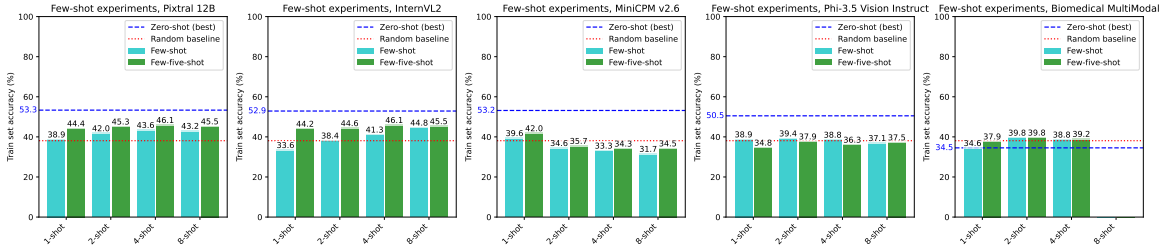


Figure 12: Visualization of VLM performance without and with additional task information in few-shot structured feature understanding

Models perform around random-guess accuracy

- General drop from best zero-shot performance baseline
- **Diminishing returns** with more examples
- Providing additional task information produces minor improvements

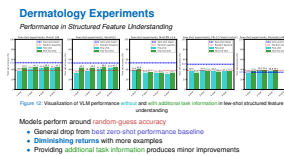
2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Evaluation

Seven-Point Checklist Dermatology Experiments

Dermatology Experiments



- Few-shot performance generally peaks at 2-4 examples
- Domain-fine-tuned model appears to have recognized the task this time, but is well in random-guessing territory like the other models

1. Introduction

1.1 Background

1.2 Problem Statement

2. Related Work

2.1 Background

2.2 Research Gaps

3. Proposed Approach

4. Methodology

4.1 Datasets

4.2 Models Chosen

4.3 Experimental Design

5. Evaluation

5.1 CIFAR-10 Experiments

5.2 Seven-Point Checklist Dermatology Experiments

6. Conclusions

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Conclusions

| | |
|---|--|
| 1. Introduction | |
| 1.1 Background | |
| 1.2 Problem Statement | |
| 2. Related Work | |
| 2.1 Background | |
| 2.2 Research Gaps | |
| 3. Proposed Approach | |
| 4. Methodology | |
| 4.1 Datasets | |
| 4.2 Models Chosen | |
| 4.3 Experimental Design | |
| 5. Evaluation | |
| 5.1 CIFAR-10 Experiments | |
| 5.2 Seven-Point Checklist Dermatology Experiments | |
| 6. Conclusions | |

Key Findings

What did we learn about VLM few-shot transfer?

General Domain (CIFAR-10)

- VLMs show **fairly good transfer** to this task
- Downstream models achieve near ground-truth performance
 - Only 5% below models trained from ground-truth
 - **Resilient** to pseudolabel noise

We find:

- Zero / low-shot approaches with **simple prompting** most effective
- Simpler prompting strategies generally outperform complex ones
- Examples help model **recognize** the task (from pre-trained priors) ⁶

⁶ J. Pan et al., What In-Context Learning “Learns” In-Context: Disentangling Task Recognition and Task Learning, in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers et al., Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8298–8319

Specialized Domain (Derm7Pt)

- **Limited transfer** to specialized domains (Derm7Pt)
 - Performance near random-guessing
 - Suggests **strong reliance** on pre-trained knowledge
- Multi-modal **integration challenges** exist

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Conclusions

Key Findings

- Zero-shot and low-shot approaches with simple prompting proved most effective for general domain tasks
- However, performance degraded significantly on specialized medical tasks (Derm7Pt). The domain is likely underrepresented in the pre-training data of the VLMs
- Architecture choices significantly impacted both performance and resource efficiency, e.g. the Perceiver-resampler architecture used in one of our models performs quite well, accuracy, memory, and time-wise
- Additional shots or reasoning requirements often increased costs without proportional gains
- Results suggest focusing on activating existing capabilities and knowledge rather than teaching new ones
- Each kind of image is a different modality in the dermatology case. The model may not know how to look at both image which are two forms of the same thing, and reason and understand the relationship between the two. The integration of the two modalities is a challenge. How do you look at to forms of the same thing and reason about the relationship between the two?

Key Findings

What did we learn about VLM few-shot transfer?

General Domain (CIFAR-10)

- VLMs show **fairly good transfer** to this task
- Downstream models achieve near ground-truth performance
 - Only 5% below models trained from ground-truth
 - **Resilient** to pseudolabel noise

Specialized Domain (Derm7Pt)

- **Limited transfer** to specialized domains (Derm7Pt)
 - Performance near random-guessing
 - Suggests **strong reliance** on pre-trained knowledge
- Multi-modal **integration challenges** exist

We find:

- Zero / low-shot approaches with **simple prompting** most effective
- Simpler prompting strategies generally outperform complex ones
- Examples help model **recognize** the task (from pre-trained priors) ⁶

⁶ J. Pan et al., What In-Context Learning “Learns” In-Context: Disentangling Task Recognition and Task Learning, in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers et al., Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8298–8319

Key Contributions

How did we help?

- Advancing the understanding of **few-shot transfer** of VLMs:
 - providing a comprehensive **evaluation framework** for assessing VLM transfer
 - providing **practical guidelines** for their transfer to dataset annotation tasks to train task-specific downstream models
- Empirical findings suggest downstream models are **resilient to label noise**:
 - significant for deploying these systems in **resource-constrained** environments
 - suggests that VLMs **can** support the learning of smaller, task-specific models despite imperfect predictions

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└─ Conclusions

└─ Key Contributions

- Advancing the understanding of **few-shot transfer** of VLMs:
 - providing a comprehensive **evaluation framework** for assessing VLM transfer
 - providing **practical guidelines** for their transfer to dataset annotation tasks to train task-specific downstream models
- Empirical findings suggest downstream models are **resilient to label noise**:
 - significant for deploying these systems in **resource-constrained** environments
 - suggests that VLMs **can** support the learning of smaller, task-specific models despite imperfect predictions

Limitations and Future Work

What constrained us, and what could be worked upon?

Technical Limitations

- Context length and memory constraints
- Limited / **degraded performance scaling** with number of examples
- **Limited generalization** to specialized domains

General Limitations

- Limited selection of models, datasets, tasks
- Findings of this work may not generalize to all vision tasks

Future Research Directions

- Generalizable **domain-transferable** pre-training strategies
- More **efficient** VLM architectures for multi-image processing
- Improving prompting strategies and **visual grounding**^a
- Extending investigation to **other vision tasks**

^aJ. Yang et al., Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. [arXiv: 2310.11441](https://arxiv.org/abs/2310.11441) [cs.CV]

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

Conclusions

Limitations and Future Work

- Context length and memory limitations restricted few-shot experiments to limited examples and constrained larger models, and high resolution images had to be rescaled to fit more examples
- Poor performance on specialized domains suggests models rely more on pre-trained priors than learning new capabilities
- There are some things that you can tell beforehand are bad ideas, and this was not specifically one of them. We expected before starting the project that more examples would strictly improve performance, but this was not the case. In NLP and LLMs, few-shot transfer and chain-of-thought prompting is extremely effective, but this is not the case for vision-language models and vision-language tasks, as we see here.
- More efficient VLM architectures needed for scalable high-resolution multi-image processing. Not only this, but for the use of VLMs and VLAs in robots in general, the ability to process multiple images at once and keep them in context is extremely important.
- Other vision tasks to try: object detection, segmentation, captioning to support a contrastive learner, etc.

Limitations and Future Work

What constrained us, and what could be worked upon?

Technical Limitations

- Context length and memory constraints
- Limited / **degraded performance scaling** with number of examples
- **Limited generalization** to specialized domains

General Limitations

- Limited selection of models, datasets, tasks
- Findings of this work may not generalize to all vision tasks

Future Research Directions

- Generalizable **domain-transferable** pre-training strategies
- More **efficient** VLM architectures for multi-image processing
- Improving prompting strategies and **visual grounding**^a
- Extending investigation to **other vision tasks**

^aJ. Yang et al., Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. [arXiv: 2310.11441](https://arxiv.org/abs/2310.11441) [cs.CV]

Thank You!

Questions?

Contact:

- shinas.shaji@smail.inf.h-brs.de
- shinas.shaji@iais.fraunhofer.de

2025-03-12

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

└─ Conclusions

└─ Thank You!

Q: Why were advanced models such as GPT-4o and other closed-source models not used in the work?

A: Multiple reasons:

- compute analysis cannot be performed, as memory usage cannot be read, times are dependent on network, and the size of the model is unknown
- we focus on open-source models, for whom the architectural details are known. This enables us to see, for example, that the perciever-resampler architecture used in one of our models performs quite well, accuracy, memory, and time-wise
- we cannot be sure that our datasets are in the dataset training mix for closed-source models. Open-source work typically also mentions the training data mix and datasets used. Remember, we needed to check how few-shot learning performs when we know that similar tasks are not in the pre-training data for the model
- closed-source models can get quite expensive when we need to push a whole dataset through them. We can however use batched inferences for half the cost. However, the code and processing would require overhauls to support this, and this was perhaps beyond scope

Thank You!

Questions?

Contact:
• shinas.shaji@smail.inf.h-brs.de
• shinas.shaji@iais.fraunhofer.de