Language as a Label: Zero-Shot Multimodal Classification of Everyday Postures under Data Scarcity

MingZe Tang, Jubal Chandy Jacob
Department of Computing Science
University of Aberdeen
{m.tang.24, j.chandyjacob.24}@abdn.ac.uk

Abstract

This paper investigates how the specificity of natural language prompts influences zero-shot classification performance in modern vision language models (VLMs) under severe data scarcity. Using a curated 285 image subset of MS COCO containing three everyday postures (sitting, standing, and walking/running), we evaluate OpenCLIP, MetaCLIP 2, and SigLIP alongside unimodal and pose-based baselines. We introduce a three tier prompt design, minimal labels, action cues, and compact geometric descriptions and systematically vary only the linguistic detail. Our results reveal a counter-intuitive trend where simpler prompts consistently outperform more detailed ones, a phenomenon we term *prompt overfitting*. Grad-CAM attribution further shows that prompt specificity shifts attention between contextual and pose-relevant regions, explaining the model dependent behaviour. The study provides a controlled analysis of prompt granularity in low resource image based posture recognition, highlights the need for careful prompt design when labels are scarce.

1 Introduction

Label scarcity is a central barrier for practical human action recognition from still images (Wu et al., 2022). Many deployments cannot acquire balanced annotations or run task specific training. Vision and language encoders mitigate this limitation by learning a shared embedding space in which text can serve as a label at inference time (Radford et al., 2021). This paper studies whether careful wording of those text labels improves zero shot classification under data scarcity.

The task focuses on three everyday postures in still images, namely sitting, standing, and walking or running, using a small subset derived from COCO (Lin et al., 2014) with 230 images. Image content, preprocessing, and scoring are held fixed, and language acts as the only supervision at inference. Each image is embedded once at the native input size of the model and is scored by cosine similarity against one prompt per class. Prompt specificity is the sole experimental factor and follows a three tier design. Tier one uses a minimal label template. Tier two adds a short action cue. Tier three adds compact pose geometry that specifies body configuration. Prompts exclude scene, identity, and clothing terms so that differences arise only from pose description.

Evaluation covers multimodal encoders that align images and text, namely OpenCLIP, MetaCLIP and SigLIP. Vision only baselines include DINOv3 and a standard Vision Transformer paired with frozen sentence embeddings to form a heuristic zero shot classifier. A pose based baseline uses YOLOv11 Pose for key-point estimation together with a simple geometric decision rule. Results are reported as accuracy and macro F1 for each tier and each model, and qualitative analysis with gradient based visualisations assesses whether greater prompt specificity shifts attention toward pose relevant regions. The study provides an empirical protocol for zero shot recognition under data scarcity and a controlled comparison of prompt wording across modern encoders and non linguistic baselines.

2 Related Works

2.1 Vision–Language Models for Zero-Shot Classification

At the core of modern zero-shot classification is the contrastive language–image pre-training paradigm introduced by CLIP, which aligns visual and textual representations in a shared embedding space using large collections of image–text pairs (Radford et al., 2021; Jia et al., 2021). The objective draws paired images and texts closer while separating mismatched pairs, thereby encoding vision–language

correspondences (Zhang et al., 2023). Zero-shot classification then becomes a nearest-neighbour search in this shared space: given an input image and a set of class descriptions, the model computes cosine similarity between image features and the text embeddings of prompts such as "a photo of a `[class]`", predicting the class with the highest similarity (Ghiasvand et al., 2025b,a).

OpenCLIP demonstrates competitive results across more than thirty benchmarks spanning OCR, scene recognition, and fine-grained object categorisation, often approaching supervised baselines without task-specific training (Radford et al., 2021). The contrastive formulation has since been adapted for downstream tasks including detection, segmentation, video action recognition, and depth estimation (Xu et al., 2023; Zhou et al., 2022; Xu et al., 2022). More recent models such as SigLIP revisit the pre-training loss, replacing the softmax contrastive objective with independent sigmoid scoring (Zhai et al., 2023), while MetaCLIP 2 improves performance by scaling and curating training data (Chuang et al., 2025). Dual-encoder VLMs remain dominant due to their scalability, robustness, and task flexibility (Volkov et al., 2025; Zhang et al., 2023).

2.2 Low-Resource & Low-Compute Image Understanding

Balancing performance and computational efficiency has motivated training-free or parameter-efficient approaches for resource-constrained environments (Zhang et al., 2024). These methods leverage pre-trained representations to extract more information from test samples and class names without parameter updates a valuable property when labelled data, compute, or both are limited. Parameter-efficient adaptation, including prompt tuning and related techniques, offers a middle ground by updating only a small subset of parameters or learned tokens while keeping backbone weights frozen (Mistretta et al., 2024; Lester et al., 2021). Training-free few-shot methods such as TIP-Adapter further exploit cached support embeddings to merge zero-shot textual knowledge with visual evidence (Esbri et al., 2024).

Resource constraints also arise from domain shift and model scale. Direct zero-shot deployment of large VLMs can be hindered by distributional mismatch, particularly in specialised settings such as medical imaging (Wang et al., 2025; Liu et al., 2023). Video understanding presents additional challenges, as spatiotemporal modeling increases computation demands (Bosetti et al., 2024; Shao et al., 2020). Collectively, these findings highlight the importance of methods that maximise generalisation under minimal supervision. VLMs function effectively as "data multipliers," transferring broad, open-world knowledge into low-resource classification settings (Zhang et al., 2024; Volkov et al., 2025). Our study takes this perspective by comparing multimodal and vision-only models under severe data scarcity.

2.3 Prompt Semantics, Granularity, and Posture Cues

Early work on domain-specific zero-shot recognition used manually specified semantic attributes to represent actions or activities, enabling recognition of unseen classes through structured linguistic or conceptual descriptions (Bosetti et al., 2024; Zellers and Choi, 2017). Later approaches replaced hand-built attributes with distributional word embeddings, framing zero-shot recognition as alignment between visual features and latent semantic spaces. These ideas extend naturally to video settings, where language acts as the primary mechanism for generalisation (Bosetti et al., 2024; Shao et al., 2020).

Prompt engineering has since become a central research focus, especially as models exhibit varying sensitivity to how class concepts are phrased. Recent studies show that prompt wording measurably affects zero-shot action recognition; prompts generated by large language models frequently outperform raw label prompts (Ali et al., 2024). Methods such as CuPL automate descriptive prompt construction and demonstrate consistent improvements on benchmarks like ImageNet (Cai et al., 2025; Pratt et al., 2022). However, Xu et al. highlight that increasing linguistic specificity does not universally improve model performance and may instead introduce distributional mismatches between text and visual cues (Xu et al., 2023).

This motivates the need to understand how multimodal encoders respond to differing levels of semantic granularity, especially for fine-grained distinctions such as human postures. Our study builds directly on this line of research: rather than generating long-form descriptions, we systematically manipulate prompt specificity minimal labels, action cues, and compact geometric wording to examine how VLMs ground pose semantics under extreme data scarcity. This controlled setup allows us to

identify a counterintuitive phenomenon we term *prompt overfitting*, where excessive detail harms performance for stronger models.

3 Data and Methods

3.1 Dataset

We evaluate on a curated 285–image subset of MS COCO Lin et al. (2014), a large–scale benchmark of everyday scenes with dense instance annotations. From the 2014 releases, images containing at least one visible person with sufficient visual evidence to judge posture were sampled, and a single action label was assigned per image by manual inspection. The subset is balanced across three classes: *sitting* (n=95), *standing* (n=92), and *walking or running* (n=98). Figure 1 shows randomly selected examples for each class and illustrates variation in viewpoint, background, and occlusion. Exploratory analysis of raw image sizes showed a concentration around 640×480 pixels, which supports a uniform resize to 224×224 for all models. Aspect–ratio distributions did not differ across classes, and visual checks confirmed negligible class–specific distortion after resizing.

Figure 1: Random samples from the curated MS COCO subset for *sitting*, *standing*, and *walking or running*.

3.2 Models and Experimental Approach

Our evaluation is structured around three distinct representation paradigms. In each case, the pre-trained model serves as a feature extractor, with a lightweight classifier trained on the resulting embeddings.

3.2.1 Unimodal Vision Models

Vision Transformer (ViT) A pre-trained ViT (Sreekanth, 2024) model (`vit-base-patch16-224`) was fine-tuned on the dataset. It was evaluated on both a binary task (sitting vs. standing) and the full three-class task to assess its generalization capability.

DinoV3 A state-of-the-art vision model pre-trained using self-supervised learning on images alone, allowing us to assess the efficacy of purely visual, non-linguistic representation learning (Siméoni et al., 2025).

3.2.2 Multimodal Vision-Language Models

OpenCLIP We employed OpenAI (Radford et al., 2021)'s CLIP pre-trained vision encoder `openai/clip-vit-base-patch32` model, leveraging its extensive image-text pre-training to generate semantically rich embeddings.

Meta CLIP 2 An evolution of CLIP, pre-trained on a more meticulously curated dataset to enhance the quality and robustness of its visual-semantic representations (**?**).

SigLip A VLM employing a sigmoid-based loss function during pre-training, offering an alternative to the contrastive objective of CLIP (**?**).

3.2.3 Pose-Centric Structural Model

YOLOv11x-pose This model implements a two-stage process. First, the YOLOv11x pose (Khanam and Hussain, 2024) architecture is applied to each image to extract a set of 2D keypoints that represents the subject's skeleton. Second, geometric features, such as the angles between the left and right knee and hip joints, are calculated from these key points. A simple classifier is then trained on these angular features to determine the final action class.

3.3 Prompt tiering for zero-shot classification

We vary only the specificity of the text prompt in order to test how wording affects zero-shot posture recognition with scarce data. **Tier 1** uses the class label in a minimal template such as "a photo of a person `[class]`", which reflects common zero-shot practice. **Tier 2** adds a brief action cue that clarifies the target category, for example "a person seated on a chair", "a person standing still and upright", or "a person mid-stride with one foot off the ground". **Tier 3** replaces action words with short anatomical or pose constraints, for example "hips and knees bent at right angles" for sitting or "legs straight and torso vertical" for standing. Across tiers we keep prompts scene free and we avoid background, clothing, and identity terms so that only pose information differs. For each tier we create one prompt per class, compute unit-normalized text embeddings once, embed each image once at the model's native resolution, and score classes by

cosine similarity. We report accuracy and macro-F1 per tier on the same images and preprocessing settings without any model fine-tuning so that observed differences can be attributed to prompt content rather than changes in data or optimization.

3.4 Experimental Approach

All experiments were conducted with a focus on reproducibility and were run on Google Colab on a single NVIDIA T4 GPU with 16GB of memory. The dataset was partitioned on a fixed stratified 80% training, 10% validation, and 10% test split for all experiments. A global random seed was established to ensure that all models were trained and evaluated on the exact same data partitions.

The task was defined in two distinct classification scenarios to assess the performance of the model at varying levels of difficulty: (1) A simplified binary task focusing on two more visually distinct classes: *sitting* vs. *walking/running*. (2) Three-class task encompassing all labels: *sitting*, *standing*, and *walking/running*; To account for stochasticity in the training process, each model was trained and evaluated over five independent runs with different seeds. We used an early stopping mechanism with patience of 5 epochs, monitoring the validation loss to prevent over-fitting.

4 Results

The empirical evaluation is presented in two parts. (1) A comparative analysis of different model architectures under a standard training and evaluation paradigm to establish baseline performance. (2) A zero-shot experiment investigating how the specificity of text prompts affects the performance of Vision-Language Models.

4.1 Comparative Analysis of Model Architectures

A comparative evaluation of models from three different paradigms (Unimodal, VLM, and Pose-Centric) was conducted across both binary and multi-class classification tasks as the first stage of the evaluation.

4.1.1 Performance on Binary Classification

To investigate how model performance is affected by task complexity, the models were evaluated on a binary classification task (sitting vs. walking/running). The results, presented in Table 1, show a general performance uplift across most models, yet the relative ranking remains largely consistent.

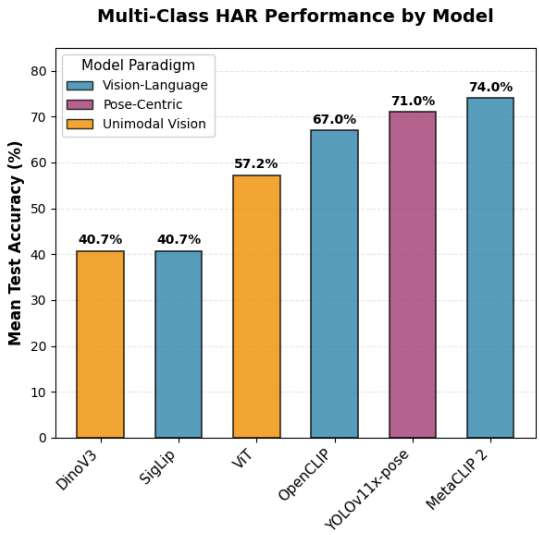

Figure 2: Mean test accuracy on the primary three-class HAR task. Models with semantic (VLM) or structural (Pose) priors demonstrate a clear performance advantage.

MetaCLIP 2 performed the best with an accuracy of 92.8%. Notably, ViT performed well in this less ambiguous setting, achieving 90.0% accuracy, nearly matching the top VLM. This suggests that when classes are more visually distinct, a powerful unimodal architecture can be highly effective. The original OpenCLIP model also performed strongly at 88.1%. However, DinoV3 and SigLip surprisingly continued to lag significantly, with accuracies of 57.5% and 56.5%, respectively.

4.1.2 Performance on Multi-Class Classification

The three-class action recognition task represents the core challenge of this study, requiring models to distinguish between visually similar and ambiguous postures from a single static frame. As detailed in Figure 2, the performance of the evaluated models diverged significantly, clearly separating them into distinct tiers.

The models endowed with strong priors formed the top tier. The Vision-Language Model MetaCLIP 2 achieved the highest accuracy at 74.0%. Following closely was the YOLOv11x-pose model, which, by leveraging a structural representation of the human body, secured an accuracy of 71.0%. The original OpenCLIP model also delivered a robust performance of 67.0%.

A substantial performance gap exists between

these models and the unimodal models that learn from pixels alone. The standard Vision Transformer (ViT) achieved a modest accuracy of 57.2%. The purely self-supervised DinoV3 and the VLM SigLip both struggled significantly, each attaining only 40.7% accuracy, a result only marginally better than random chance.

Model	B. Acc.	M. Acc.	Prec.	Rec.	F1
MetaCLIP 2	0.92	0.74	0.74	0.74	0.74
ViT	0.90	0.52	0.59	0.57	0.57
OpenCLIP	0.88	0.67	0.68	0.67	0.66
YOLOv11x-pose	—	0.71	0.73	0.71	0.71
DinoV3	0.57	0.40	0.41	0.41	0.40
SigLip	0.56	0.40	0.28	0.41	0.33

Table 1: Performance comparison of vision models, reporting **(B. Acc.)** Binary Accuracy, **(M. Acc.)** Multi-class Accuracy, **(Prec.)** Macro Precision, **(Rec.)** Macro Recall, and **(F1)** Macro F1 Score.

4.1.3 Analysis of Class-Specific Metrics

To gain a more nuanced understanding, we analyzed the macro-averaged Precision, Recall, and F1-Score (Table 1). These metrics reinforce the hierarchy observed in accuracy. MetaCLIP 2 and YOLOv11x-pose demonstrated a strong balance between precision and recall, resulting in high F1-Scores of 0.74 and 0.71, respectively, indicating reliable classification across all three categories. In contrast, lower-performing models exhibited imbalances. For instance, SigLip2 had a recall of 0.41 but a very low precision of 0.28, suggesting it generated a large number of false positive predictions in its attempt to classify instances from all classes.

4.2 Prompt-Specific Zero-Shot Performance

In our second set of experiments, we investigated how prompt specificity affects the zero-shot performance of VLMs. The results, detailed in Table 2, show that the relationship between prompt detail and model performance is not linear and is highly model-dependent.

4.2.1 Performance Trends for High Performing Models

The primary trend observed for the leading Vision-Language Models is a clear inverse relationship between prompt specificity and classification performance. As represented in Table 2 both MetaCLIP 2 and OpenCLIP, the simplest Tier 1 prompts consistently achieved the highest accuracy and F1 scores. The introduction of more descriptive features in Tier 2 or anatomical cues in Tier 3 resulted in a significant degradation of performance. This effect was particularly pronounced for MetaCLIP 2, where the multi-class accuracy fell sharply from 68.8% with a Tier 1 prompt to 55.1% with a Tier 2 prompt. Similarly, OpenCLIP's multi-class accuracy saw a substantial decrease from a high of 71.2% (Tier 1) to 52.6% (Tier 2). This consistent impact suggests a phenomenon of "prompt overfitting" where excessive detail may unduly constrain the models and hinder their ability to generalize.

4.2.2 Model-Dependent Responses to Prompt Granularity

The Tier 1 performance trend was not universal, highlighting that the optimal prompt strategy is highly model dependent. The lower-performing SigLip model exhibited a contrasting response to the increase in prompt detail. While its overall accuracy remained consistently low, its ability to classify the ambiguous `walking_running` class was significantly boosted by the specific, "body cue-based" Tier 3 prompts. This was most evident in the binary task, where the F1 score for this specific class jumped from 0.364 with a basic Tier 1 prompt to 0.566 with the detailed Tier 3 prompt.

5 Discussion & Conclusion

5.1 Prompt Specificity as Supervision at Inference

Prompt wording functions as an explicit prior on the classifier decision in a zero-shot setting. Minimal, noun-centric prompts align with the distributions seen during pre-training of image–text encoders, where concept names are frequent and broadly grounded. This alignment explains the competitiveness of label-only prompts in closed sets. In contrast, adding brief action cues can introduce a linguistic–visual mismatch for still images, since verbs such as "walking" or "standing still" denote dynamics or intent rather than stable appearance. The resulting text embeddings are drawn toward contexts that are weakly supported by a single frame, which reduces similarity margins and increases overlap between neighbouring classes.

Geometric phrasing exerts a different influence. Short anatomical constraints specify local, view-stable relations, for example relative angles at the hip and knee or verticality of the torso, that are directly verifiable in a single image. Gradient-based attributions (see Figure 3a) consistently show increased concentration over limb and torso regions

Task	Tier	MetaCLIP 2		OpenCLIP		SigLIP	
		Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Binary	Tier 1	0.938	0.938	0.907	0.907	0.565	0.516
	Tier 2	0.751	0.742	0.850	0.847	0.523	0.490
	Tier 3	0.731	0.714	0.876	0.875	0.539	0.537
Multi-class	Tier 1	0.688	0.686	0.712	0.708	0.365	0.346
	Tier 2	0.551	0.508	0.526	0.533	0.316	0.259
	Tier 3	0.565	0.528	0.628	0.628	0.312	0.302

Table 2: Classification Performance (Binary and Multi-class) across different tiers of prompts

(a) Grad-CAM for three phrasings of the "sitting" concept, showing saliency on the chair and hip–knee region.

(b) Grad-CAM for three phrasings of the "standing" concept, showing saliency on the legs and torso.

Figure 3: Grad-CAM visualizations for different phrasings of the concepts "sitting" and "standing". Increased specificity in phrasing leads to more focused saliency in relevant body regions.

when such constraints are used, and decreased reliance on background texture or incidental objects. The benefit is class dependent. Categories that are well captured by a nominal phrase, such as sitting in uncluttered scenes, receive limited additional gain. Categories that are visually adjacent in a still image, such as standing versus walking or running, benefit from geometric prompts because these encode spatial structure that separates the classes without introducing scene bias.

These observations support a simple policy for low-resource use. Prefer label-style prompts as the default in closed-set classification with pre-trained encoders. Introduce compact geometric descriptors selectively for pairs that remain ambiguous, and verify with attribution that attention shifts from background to pose-relevant regions. Reserve action verbs for cases where the class definition truly requires dynamic semantics, since such wording is not consistently grounded in single images.

5.2 Comparative Model Behaviour and Calibration

Across encoders, closed-set zero-shot performance tracks the pre-training objective. OpenCLIP and MetaCLIP optimise a soft-max contrastive loss with a learned temperature, which induces competition among text candidates and yields larger similarity margins in classification. SigLIP optimises independent sigmoid scores for pairs, which favours retrieval but produces flatter score distributions in a closed set. The flatter distributions manifest as smaller top-1 minus top-2 margins and greater sensitivity to prompt phrasing, particularly for visually adjacent classes.

Calibration follows the same pattern. After unit normalisation of embeddings, a single temperature applied to cosine similarities brings CLIP-

family confidences into closer agreement with accuracy. The same treatment is less effective for SigLIP model because the training objective does not enforce cross-class competition, and confidence therefore reflects pairwise affinity rather than calibrated class probability. Reliability curves and expected calibration error consequently favour OpenCLIP and MetaCLIP 2 under a shared temperature, whereas SigLIP remains comparatively miscalibrated or require tier-specific scaling.

Baselines clarify the role of alignment and structure. DINOv3 and a standard ViT combined with frozen sentence embeddings underperform and calibrate poorly because the image and text spaces are learned independently rather than jointly. YOLOv11-Pose with a simple geometric decision rule is competitive when keypoints are detected with confidence, which indicates that explicit pose structure can substitute for language supervision when the detector is reliable. Taken together, these observations suggest that, in data-scarce image-based recognition, cross-modal alignment with a contrastive objective provides stronger closed-set behaviour, while geometric priors provide a complementary path when alignment is weak or text supervision is constrained.

5.3 Language-Free and Pose-Based Baselines under Data Scarcity

Vision-only encoders such as DINOv3 and a standard ViT provide a language-free reference that isolates the value of cross-modal alignment. When image embeddings are compared to sentence embeddings from an unrelated text model, the spaces are not jointly learned. As a result the cosine geometry reflects two independent objectives rather than class evidence. This mismatch explains the weaker separability and the poor calibration that appear even when preprocessing is held constant. The baselines are therefore informative as a lower bound. They confirm that generic visual features carry some signal for posture, yet they also show that alignment with text during pre-training is the primary driver of robust zero-shot classification.

A pose-based baseline introduces a different kind of supervision that is structural rather than linguistic. YOLOv11-Pose produces 2D keypoints, and a deterministic rule maps joint configuration to the three classes. When detections are confident, the rules are competitive because they test explicit geometric relations that are stable in a single frame. However, performance depends on detection coverage. Occlusion, truncation, unusual viewpoints, and multiple persons reduce keypoint quality and lead to abstentions or incorrect geometry, which directly lowers accuracy. Reporting coverage alongside accuracy is therefore necessary. On the covered subset the baseline demonstrates that posture can be resolved without any text, while the uncovered subset clarifies where structural priors fail.

These baselines contribute two practical insights for low-resource use. First, if language supervision is restricted because of privacy or deployment constraints, a pose pipeline can recover a substantial fraction of performance provided that person detection is dependable. Second, if a language-free heuristic is required for simplicity, cosine scoring between DINOv3 or ViT features and frozen sentence embeddings should be treated as a diagnostic tool rather than as a calibrated classifier. In combination with the multimodal results, the baselines indicate that cross-modal alignment should be the default, and that explicit pose structure is a useful fallback when alignment is unavailable or when prompts cannot be used.

5.4 Attention Maps and Error Patterns in Still-Image HAR

Attribution on the sitting examples shows a consistent shift as prompt specificity increases. The label-style prompt yields broad responses that cover the person and nearby objects. Adding an action cue narrows the response toward the pelvis and the supporting surface. Geometric phrasing concentrates the map on hips, knees, and the contact region with the chair. This progression indicates that geometric wording encourages the model to prefer pose evidence over contextual cues.

For standing, the label-style prompt again produces diffuse maps with noticeable activation on salient background regions as seen in Figure 3b. The action cue that mentions stillness reduces spread and increases activation around the legs and feet. The geometric formulation further localises energy along the vertical axis of the body, especially the shins and torso. When predictions are incorrect for standing, the maps typically remain broad and include background structure, which suggests insufficient reliance on limb configuration in those cases.

Two practical uses follow. First, attribution can serve as a prompt diagnostic: adopt geometric phrasing when maps remain diffuse under a label-

style prompt, and retain the minimal prompt when maps are already concentrated on limbs and joints. Second, report simple map statistics alongside accuracy, such as the proportion of normalised heat inside a person region and the entropy of the map. Higher in-person proportion and lower entropy correlate with the tighter, pose-focused responses observed for the geometric prompts in both sitting and standing.

5.5 Practical Implications, Robustness, and Limitations

Practical implications In data-scarce settings, a label-style prompt for each class with unit-normalised embeddings and a single temperature applied to cosine scores is a strong baseline. When confusions remain for visually adjacent categories, replace the label with a compact geometric description for those specific classes. Monitor decision confidence with the top-one minus top-two similarity margin. Introduce an abstention rule based on a margin threshold for low-confidence cases. If language supervision is not available, a pose pipeline that uses YOLOv11-Pose with a deterministic geometric rule provides an alternative, provided that keypoint detection is reliable.

Robustness considerations Performance depends on image framing and resolution. Crops that remove feet or hips reduce margins for posture classes, therefore detection and resizing should preserve the lower body. Resolution influences prompts that encode limb configuration. The native 224 input supports fair comparison, while higher resolution can improve separation when resources allow. Paraphrases within a tier can shift scores, so a small prompt ensemble per class stabilises predictions with limited overhead. Calibration differs across encoders. Fit a single temperature once per model and keep it fixed across tiers to preserve comparability. For the pose baseline, report coverage since occlusion, truncation, and small subjects reduce the fraction of usable detections.

Limitations Utilizing a small COCO-derived subset (285 images) focusing only on three single-person posture classes (sitting, standing, walking/running) from still images, inherently limits its external validity and applicability to real-world Human Activity Recognition (HAR) where temporal cues are essential. Additionally, the discovery of "prompt overfitting" is based solely on hand-crafted English prompts across a limited set of encoders (OpenCLIP, MetaCLIP 2, SigLip) and verified using only Grad-CAM for qualitative analysis. This means the conclusions apply most directly to static, image-based posture classification under severe data scarcity.

6 Conclusion

This study examined how natural language prompt specificity functions as a form of supervision at inference time for zero shot posture recognition under extreme data scarcity. By holding all visual processing constant and varying only the wording of class prompts across three tiers, we identified a consistent and counterintuitive pattern in modern VLMs: stronger encoders such as MetaCLIP 2 and OpenCLIP perform best with minimal label-style prompts, while additional descriptive detail reduces accuracy, a phenomenon we term *prompt overfitting*. Conversely, lower-performing models benefit from compact geometric descriptions, particularly for visually adjacent postures. Attribution analyses further showed that prompt wording shifts model attention toward or away from pose relevant regions, clarifying why specificity can help or hinder depending on encoder strength.

Together with comparisons to vision only and pose based baselines, these findings provide practical guidelines for deploying VLMs in low-resource, image-based Human Activity Recognition (HAR) settings. Label-style prompts serve as a strong default, while geometric descriptions can assist when distinctions are subtle and model capacity is limited. More broadly, the results demonstrate that prompt semantics act as implicit priors that can either reinforce or misalign the visual grounding of multimodal encoders.

Future work should extend this analysis to larger and more diverse datasets, multi-person and occlusion-heavy scenes, multilingual and LLM-generated prompts, additional VLM backbones, and more advanced attribution techniques. Such work is necessary to fully understand and generalise the behavioural patterns observed in this study.

References

Mahmoud Ali, Di Yang, and Franccois Br'emond. 2024. Are visual-language models effective in action recognition? a comparative study. *ArXiv*, abs/2410.17149.

Massimo Bosetti, Shibingfeng Zhang, Bendetta Liberatori, Giacomo Zara, Elisa Ricci, and Paolo Rota.

2024. Text-enhanced zero-shot action recognition: A training-free approach. *ArXiv*, abs/2408.16412.

Lincan Cai, Jingxuan Kang, Shuang Li, Wenxuan Ma, Binhui Xie, Zhida Qin, and Jian Liang. 2025. From local details to global context: Advancing vision-language models with attention-based selection. *ArXiv*, abs/2505.13233.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. 2025. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2501.01234*.

Pablo Meseguer Esbri, Rocío del Amor, and Valery Naranjo. 2024. Mi-visionshot: Few-shot adaptation of vision-language models for slide-level classification of histopathological images. *ArXiv*, abs/2410.15881.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. 2025a. pfedmma: Personalized federated fine-tuning with multi-modal adapter for vision-language models. *ArXiv*, abs/2507.05394.

Sajjad Ghiasvand, Haniyeh Ehsani Oskouie, Mahnoosh Alizadeh, and Ramtin Pedarsani. 2025b. Few-shot adversarial low-rank fine-tuning of vision-language models. *ArXiv*, abs/2505.15130.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, Deva Ramanan, Piotr Dollár, and C. L. Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Marco Mistretta, Alberto Baldrati, Marco Bertini, and Andrew D. Bagdanov. 2024. Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation. In *European Conference on Computer Vision*.

Sarah Pratt, Rosanne Liu, and Ali Farhadi. 2022. What does a platypus look like? generating customized prompts for zero-shot image classification. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15645–15655.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Hao Shao, Shengju Qian, and Yu Liu. 2020. Temporal interlacing network. *ArXiv*, abs/2001.06499.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. Dinov3.

Sreekanth. 2024. Sreekanth3096/vit-coco-image-classification · hugging face. *Sreekanth3096/vit-coco-image-classification · Hugging Face*.

Illia Volkov, Nikita Kisel, Klára Janousková, and Jirí Matas. 2025. Image recognition with vision and language embeddings of vlms.

Qian-Wei Wang, Yuqiu Xie, Letian Zhang, Zimo Liu, and Shu-Tao Xia. 2025. Pre-trained vision-language models assisted noisy partial label learning. *ArXiv*, abs/2506.03229.

Ming-Kuan Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, Xiaoshuai Sun, and Rongrong Ji. 2022. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *AAAI Conference on Artificial Intelligence*.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and X. Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18123.

Zhenlin Xu, Yi Zhu, Tiffany Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. 2023. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1827–1836.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. *ArXiv*, abs/1707.09468.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5625–5644.

Yabin Zhang, Wen-Qing Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. 2024. Dual memory networks: A versatile adaptation approach for vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28718–28728.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krahenbuhl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. *ArXiv*, abs/2201.02605.