Vision-Language Modeling for Natural-Language Wheel Loader Assistance in Unstructured Construction Environments

Kumar Manas

***Abstract*— Autonomous operation of heavy construction machinery in unstructured outdoor environments remains a critical open challenge in field robotics. Construction sites present a compound set of difficulties, including variable terrain, dust, occlusion, and continuously evolving material layouts, that resist the assumptions underlying most autonomous systems. A further fundamental challenge is that operator intent is expressed in natural language rather than precise waypoints, demanding tight integration of perception, language understanding, and action generation. We present, as a *system description* paper, a hybrid vision-language modeling framework for natural-language-guided wheel loader assistance. The system is VLA-*inspired* in its overall information flow (vision + language $\rightarrow$ action symbol), but is explicitly *not* an end-to-end visuomotor policy: the output is a structured, image-plane action string decoded by a rule-based parser rather than a continuous control signal. The system combines a cascaded hybrid perception module (closed-vocabulary detection with CLIP re-ranking, open-vocabulary OWL-ViT fallback, and chromatic segmentation), automated grounded VQA corpus construction from unannotated point-of-view footage, and parameter-efficient LoRA fine-tuning of a compact language model for spatially grounded action prediction. A deterministic parsing layer with hierarchical fallback guarantees valid action output under adverse perception conditions. Initial qualitative evaluation demonstrates that the proposed hybrid design produces correctly structured, spatially grounded action sequences across primary wheel loader operational modes (approach, excavation, loading, and discharge), and maintains valid outputs under sparse training data and noisy detections. We analyse the system's current limitations and identify the key open problems: metric 3D grounding, closed-loop execution, calibrated uncertainty, and quantitative benchmarking with ground-truth action data, which must be addressed to advance towards deployable field autonomy.**

I. Introduction

The automation of heavy machinery in construction environments represents a domain where robotic capability is urgently needed yet remarkably underdeveloped relative to progress in controlled-environment robotics. Construction sites account for a substantial share of industrial fatalities in excavation and earth-moving contexts, with a significant fraction of incidents involving the operation of earth-moving and loading equipment in cluttered, dynamic, and hazardous conditions [24]. Beyond safety, autonomous construction machinery promises transformative gains in operational efficiency, enabling 24-hour operation, consistent performance under adverse environmental conditions, and remote deployment in areas too dangerous or remote for sustained human presence. A representative wheel-loader operation in an unstructured site environment is illustrated in Fig. 1, which captures the kind of visual, terrain, and material-layout variability that motivates our natural-language-guided assistance approach.

Fig. 1. Wheel loader operation in an unstructured construction setting carrying piles of construction material. Such scenes illustrate the perception and decision-making challenges motivating natural-language-guided assistance.

Yet the construction sector remains overwhelmingly reliant on human operators. The central reason is that construction environments resist the assumptions that underpin most autonomous systems: static maps, predictable object categories, structured workspaces, and the possibility of exhaustive pre-deployment programming. A wheel loader engaged in typical site operations must localise and approach material piles whose position, shape, and composition change continuously throughout a working day; adapt to muddy, uneven terrain that invalidates pre-built maps; respond to instructions delivered as natural language rather than precise coordinate commands; and operate safely in the proximity of other vehicles and personnel – often under poor lighting, heavy dust, or adverse weather.

These properties collectively define the challenge of *long-term autonomy in the wild*: robots must be persistent, adaptive, and interpretable over extended deployments in environments that resist clean formalisation [1]. Addressing this challenge requires tight integration of robust perception that tolerates appearance change, semantic understanding of natural language operator intent, and reliable translation of high-level goals into physically executable behaviours – precisely the capabilities that Vision-Language-Action (VLA)

Kumar Manas is with the Interactive Robot Perception & Learning (PEARL) Group, Technical University of Darmstadt, Germany. kumar.manas@tu-darmstadt.de

Part of this work was carried out while the author was with Freie Universität Berlin, Berlin, Germany.

models aspire to provide.

Recent VLA research [2]–[4] has demonstrated that jointly training over visual, linguistic, and action modalities yields remarkable generalisation in tabletop manipulation. The core insight is compelling: rather than engineering task-specific control pipelines, a single model pre-trained on the rich semantic structure of language and vision can acquire broad behavioural competence. However, existing VLA work is almost exclusively evaluated in indoor manipulation settings. The extension to outdoor heavy machinery involves not merely a change of scene, but a fundamental shift in the nature of every component: perception must handle open-set object categories under extreme appearance variation; language commands are goal-directed rather than step-directed; and the consequences of failure extend beyond a dropped object to equipment damage or personnel injury.

Crucially, the data acquisition problem is qualitatively different in the field. Large-scale demonstration datasets – the fuel of modern VLA training – are straightforward to collect in laboratory manipulation settings. On construction sites, collecting paired demonstrations with labelled actions requires significant operational disruption. This motivates a core design principle of our work: the training corpus must be constructable from freely available, unannotated operational footage, without requiring expert annotation or controlled data collection campaigns.

Scope and positioning. We position this paper as a *system description* of a prototype hybrid vision-language pipeline rather than as a benchmarked VLA model. Our system is *VLA-inspired* in that it follows the vision-plus-language-to-action information flow, but it differs from true VLA models [2], [4] in two important respects: (i) the action "space" consists of structured natural-language strings parameterised by normalised image-plane coordinates, not low-level motor commands or continuous control signals; and (ii) the final mapping from generated text to a valid action is performed by a deterministic rule-based parser, not by a learned policy head. The role of the language model is therefore to provide *semantic grounding and command interpretation*, while the parser provides syntactic guarantees needed for downstream control interfaces. We make this distinction explicit throughout the paper to avoid overstating the system's relationship to end-to-end VLA literature.

This paper presents a hybrid VLM-plus-action-parsing system for construction wheel loader assistance that embodies this principle. Our contributions are:

1) A **zero-annotation data generation pipeline** that builds a construction-domain training corpus from unannotated point-of-view footage via open-vocabulary detection and vision-language captioning.
2) A **robust hybrid inference stack** that couples cascaded perception (closed-vocabulary detection with open-vocabulary and chromatic fallbacks) with deterministic action parsing to preserve valid, spatially grounded outputs under noisy scene conditions.
3) A **compute-efficient and reproducible implementation** that combines LoRA-based adaptation of a compact instruction model with an open interactive demo for straightforward community evaluation and extension.

II. Related Work

A. Long-Term Autonomy in Unstructured Environments

The challenge of persistent robot operation in the wild has been studied across outdoor navigation, environmental monitoring, and planetary exploration. Cadena *et al.* [15] surveyed long-term simultaneous localisation and mapping, identifying appearance change, dynamic objects, and scene evolution as fundamental obstacles that static map representations cannot address. Doherty *et al.* [16] argued for probabilistic environmental representations that model uncertainty explicitly to enable safe long-horizon planning as scene statistics drift [17]. In field robotics more broadly, systems operating in unstructured terrain must contend with changing contact conditions and mobility constraints, motivating continual adaptation strategies [18]. The construction domain presents analogous challenges with the added complexity that the material being processed actively reshapes the environment throughout each operational shift – material piles are consumed, relocated, and created as work proceeds, requiring perception systems that are both reactive and temporally consistent.

B. Vision-Language-Action Models

The integration of large language models with visual perception and robot control has advanced rapidly. SayCan [13] demonstrated that language model priors can be grounded in robot affordances for long-horizon task planning, though relying on separate, pre-trained skill primitives. Code-as-Policies [14] generated executable control programs directly from natural language, treating the language model as a policy programmer. RT-2 [2] represented robot actions as language tokens within a jointly trained vision-language model, enabling knowledge transfer from web-scale pre-training to robot control. OpenVLA [4] open-sourced a generalised manipulation policy demonstrating strong cross-task generalisation. UniVLA [**?**] extended the VLA paradigm toward cross-embodiment policies through latent action representation, suggesting that the approach may eventually unify diverse robot morphologies.

In contrast to these end-to-end VLA models, which jointly learn perception, language understanding, and low-level control, our system is deliberately modular: language-conditioned semantic reasoning is performed by an adapted language model, while action validity and parameterisation are enforced by explicit symbolic post-processing. We adopt this design because it permits training on a small, automatically generated corpus and provides hard guarantees on action well-formedness that purely learned policies cannot currently match. Essentially most prior VLA evaluation is conducted in indoor manipulation settings with controlled illumination and predictable object sets. The construction domain presents a distinct and more challenging instantiation of the VLA problem that remains underexplored in the current literature.

C. Perception for Construction and Field Environments

Object detection in construction scenes is challenging because of the irregular geometry, variable texture, and extreme appearance diversity of construction materials and equipment. Fang *et al.* [21] illustrated both the feasibility and practical challenges of deploying computer-vision safety monitoring in construction imagery. The emergence of contrastive vision-language models [6] has opened the possibility of zero-shot object recognition without curated class labels, addressing the open-set nature of construction materials. OWL-ViT [7] extended this to open-vocabulary bounding-box prediction conditioned on free-form text queries at test time – a natural fit for domains where object categories are not exhaustively known in advance. Modern real-time detection frameworks [8] provide the throughput needed for on-board deployment on construction vehicles. Our work combines all three detection paradigms in a principled cascade, exploiting each method's distinct strengths across the diverse appearance conditions encountered in the field.

D. Automated Annotation and Data Generation

The reliance on large manually annotated datasets is a persistent bottleneck for field robotics. Simulation-to-real transfer [23] partially addresses this but introduces a domain gap that must be explicitly bridged. More recently, vision-language foundation models have been used to automatically generate training annotations from raw images [9], [10], [20], supporting large-scale weakly supervised annotation workflows. Our VQA generation pipeline follows this philosophy, treating freely available operational footage as an unsupervised source of domain knowledge. To our knowledge, this is the first application of automatic annotation to construction robot action prediction.

E. Parameter-Efficient Fine-Tuning

Adapting large pre-trained language models to specialised domains without catastrophic forgetting or prohibitive compute is central to deploying such models outside well-resourced research settings. Low-Rank Adaptation (LoRA) [12] constrains weight updates to a low-rank factored form, reducing trainable parameters by orders of magnitude while preserving most of the pre-trained model's representational capacity. This approach is particularly well-suited to our setting, where the domain shift from general instruction-following to construction action prediction is substantial but the available training data is limited.

III. System Architecture

The proposed system is structured as a four-stage pipeline: automated corpus construction from unannotated footage, grounded VQA generation, language model adaptation, and structured inference with hierarchical fallback. Fig. 2 provides an architectural overview.

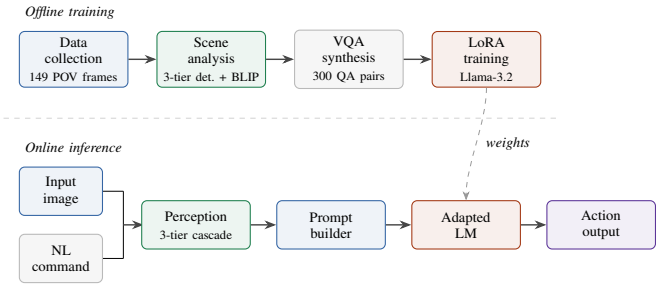

Fig. 2. VLA-inspired architecture overview. **Top (offline training)**: POV footage is processed by a scene analysis module (cascaded three-tier detector with BLIP captioner) to automatically generate VQA training pairs; these fine-tune LLaMA-3.2-1B via LoRA. **Bottom (online inference)**: a construction image and natural-language operator command are merged, processed by the same perception module, and formatted into a structured prompt for the adapted language model; its output is decoded into a grounded robot action via a rule-based fallback parser.

A. Automated Corpus Construction

A core design principle is that training data must be obtainable without manual annotation. Publicly available point-of-view footage from wheel loader operations provides a rich and accessible source of domain-relevant visual content, capturing realistic illumination conditions, motion dynamics, dust, and partial occlusion as experienced by an on-board perception system. For this work, frames were sampled from three publicly available wheel-loader point-of-view recordings hosted on YouTube, which depict typical loading, carrying, and dumping operations on open construction and quarry sites; the exact source URLs are documented alongside the data-processing scripts in our public code repository.[1] Frames are sampled at uniform temporal intervals, targeting sufficient coverage of each operational phase – approach, loading, carry, and deposit – while avoiding near-duplicate frames that would inflate dataset size without adding informational content.

B. Cascaded Hybrid Perception

Reliable spatial grounding of language model outputs requires robust detection of construction-relevant objects. No single detection paradigm achieves this reliably across the full range of outdoor construction conditions, motivating a cascaded approach in which three complementary methods are applied in order of decreasing computational cost and increasing robustness:

Tier 1: Semantic Re-classification of Region Proposals. A real-time object detector provides class-agnostic region proposals at high throughput. Each proposal is scored by a contrastive vision-language encoder [6] against a curated vocabulary of domain-relevant descriptions covering material types (gravel, sand, loose earth), construction equipment categories, and excavation zones. This two-stage design combines the computational efficiency of learned region proposals with the open-vocabulary flexibility of image-text contrastive models.

Tier 2: Open-Vocabulary Joint Detection (Fallback). When the first tier yields insufficient object coverage, an

[1] https://github.com/kumarmanas/Robot_VLA

open-vocabulary detection model [7] is invoked. This model performs detection and semantic classification jointly conditioned on free-form text queries, providing higher recall for unusual object configurations or appearance conditions not well-handled by the proposal re-classification approach.

Tier 3: Chromatic Region Segmentation. As a lightweight safety net, colour-space thresholding identifies regions whose chromatic profile is consistent with exposed earth and rock material or heavy construction equipment, a signal that remains informative when structural detectors fail due to severe occlusion or non-standard perspective. Contiguous regions above a minimum spatial extent are treated as candidate detections.

All detected objects are represented as normalised image-plane bounding boxes and centroids, enabling coordinate-conditioned action generation that is invariant to image resolution and camera parameters. In frames where construction equipment is detected but no material piles are found – a common occurrence during transit phases – synthetically positioned pile candidates with conservatively reduced confidence scores are introduced to maintain training data coverage across all operational phases.

C. Grounded VQA Generation

For each extracted frame, a vision-language captioning model [9] provides a natural-language description of the visual scene content. This caption, combined with the structured detection output, forms the basis for automatic synthesis of grounded question-answer pairs.

Questions are drawn from a template library designed to span the operational vocabulary of wheel loader control, including localisation queries (determining the position of the nearest material pile), state assessment queries (identifying suitable excavation zones), and action-selection queries covering both single-step decisions and composite operational goals. The library explicitly includes the kind of high-level, goal-directed language that operators naturally use – e.g., instructions to execute a complete loading or dumping cycle – which require decomposition into ordered primitive sequences.

Answers are generated by a deterministic resolver that maps each query type and its associated detection context to a structured element of the robot action space:

$$\begin{aligned}\mathcal{A} = \{&\textit{navigate}(\mathbf{p}),\ \textit{excavate}(\mathbf{p}),\\ &\textit{lift}(),\ \textit{discharge}(),\ \textit{turn}(\theta),\ \ldots\}\end{aligned} \tag{1}$$

where $\mathbf{p} \in [0,1]^2$ is a normalised image-plane target coordinate derived from the centroid of the relevant detected object, and θ specifies a heading adjustment. We emphasise that this action space is symbolic and image-plane: actions are textual strings parameterised by 2D image coordinates, not metric 3D waypoints or motor commands. The confidence score of the associated detection modulates action selection: high-confidence detections yield direct navigation actions, while lower-confidence detections elicit cautious approach behaviours. Composite commands produce semicolon-delimited ordered action sequences corresponding to full operational cycles such as loading (navigate $\rightarrow$ lower implement[2] $\rightarrow$ excavate $\rightarrow$ raise implement) or discharging (navigate to discharge[3] zone $\rightarrow$ position for discharge $\rightarrow$ actuate discharge $\rightarrow$ reverse clear).

D. Language Model Adaptation

The VQA corpus is formatted as structured prompts that concatenate the visual scene description (from captioning and detection serialisation), the operator command, and the target action. A general-purpose instruction-following language model [11] is fine-tuned to predict the action token sequence given this structured context.

Low-Rank Adaptation [12] is applied to all self-attention and feed-forward projection matrices, introducing trainable rank-r weight perturbations while keeping the pre-trained parameters frozen. With rank $r = 8$ and scaling factor $\alpha = 32$, the additional trainable parameter count is less than 0.5% of the full model size. This design choice is deliberate: it makes the fine-tuning step accessible on a single commodity GPU within a practical time budget, lowers the barrier for construction technology practitioners to adapt the system to their specific operational context, and preserves the general linguistic and instruction-following capabilities of the base model that are needed to handle the diversity of real operator language.

E. Structured Inference and Hierarchical Fallback

At inference time, the perception module is applied to the input image and the resulting scene representation – pile count, centroid coordinates, detection confidence, vehicle presence – is serialised into a natural-language context block prepended to the operator command. The adapted model generates an action string from this combined prompt.

A three-level post-generation parser extracts the intended action. The first level identifies composite command patterns that correspond to known multi-step operational sequences, generating the complete ordered sequence deterministically without requiring the language model to explicitly enumerate each step. The second level applies pattern matching to extract structured single-step actions and their spatial parameters, clamping all coordinate outputs to the normalised image plane. The third level provides keyword-based rule extraction as a safety guarantee, ensuring a valid action is always produced regardless of model output quality. This layered design separates semantic reasoning from syntactic robustness, an important property for deployment in environments where model failures must not translate into system failures.

IV. Experiments

We frame the experimental study below as a *system-level qualitative demonstration* of the prototype, consistent with the scope statement in Section I. Quantitative perception and action-prediction benchmarks are deferred to future work,

[2]Here, *implement* refers to the wheel loader's working attachment, typically the front bucket used for scooping, lifting, and dumping material.

[3]Here, *discharge* means unloading or releasing the carried material from the bucket at the target dumping location.

where they will be coupled with paired ground-truth operator demonstrations (see Sec. V).

A. Dataset and Training Configuration

Our current corpus is derived from three publicly available point-of-view wheel-loader video sequences processed through our automated pipeline. This yields 149 extracted frames and 300 generated question-action pairs; Table I summarises the resulting dataset and training setup.

For adaptation, we fine-tune a compact instruction model (Llama-3.2-1B-Instruct) using LoRA (rank 8, α=32, dropout 0.1), with learning rate 10^{-4}, mini-batch size 4, gradient accumulation of 2 (effective batch size 8), and three training epochs. This configuration is intentionally lightweight and serves as a practical baseline for preliminary investigation rather than deployment-ready performance. In future work, we plan to evaluate larger (e.g., 7B-parameter) foundation models as we scale our work.

TABLE I

DATASET AND TRAINING SUMMARY

Property	Value
Source video sequences	3
Extracted frames (`frame_paths`)	149
Generated training pairs (`vqa_pairs`)	300
Question templates observed	8
Action primitive types observed	10
Default detection threshold	0.4
Base model	Llama-3.2-1B-Instruct
Fine-tuning method	LoRA (r=8, α=32)
Epochs / effective batch size	3 / 8

The empirical distribution of generated question templates and action primitives is shown in Fig. 3, illustrating the coverage profile of the automatically constructed training corpus.

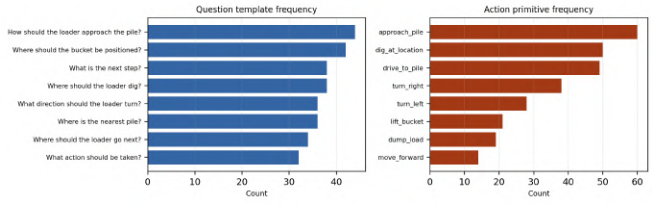

Fig. 3. Distributions computed from the generated `vqa_pairs.json` corpus (300 pairs): question-template frequency (left) and action-primitive frequency (right).

B. Perception Module Analysis

Table II summarises the complementary operational properties of the three detection tiers. Each method addresses distinct failure modes: the semantic re-classification tier handles the majority of detections in clear, well-lit scenes efficiently; the open-vocabulary tier recovers detections in partially occluded or unusual-perspective frames where rigid classification fails; and chromatic segmentation provides reliable coverage in high-dust and adverse-colour conditions at negligible computational cost. The cascade terminates as soon as sufficient detections are obtained, minimising average inference time. We note that the entries in Table II are qualitative design summaries – per-tier precision/recall and latency measurements on a labelled in-domain test set are an explicit item of ongoing work, requiring frame-level annotation that the present zero-annotation pipeline deliberately avoids.

TABLE II

PERCEPTION CASCADE: COMPLEMENTARY OPERATIONAL PROPERTIES (QUALITATIVE DESIGN SUMMARY)

Tier	Latency	Vocab.	Occl.	Role
Semantic re-cls.	Low	Closed	Moderate	Primary
Open-vocab det.	Medium	Open	High	Fallback
Chromatic seg.	Very low	Colour	Low	Safety net

C. Action Generation Evaluation

Table III presents representative system outputs for commands spanning the primary operational modes of wheel loader control. These results are qualitative and illustrative; a quantitative held-out action-prediction accuracy study, requiring ground-truth operator action labels, is left as future work. In all the cases shown, the generated actions are spatially grounded in the perceived scene: navigation targets correspond to detected material pile centroids; composite commands produce complete, correctly ordered operational sequences; and commands issued in scenes without detectable piles yield plausible conservative default behaviours.

TABLE III

REPRESENTATIVE ACTION GENERATION RESULTS (QUALITATIVE EXAMPLES)

Operator Command	System Output
"Move to the material pile"	Navigate to detected pile centroid at image position $(0.62,\ 0.57)$
"Where should I dig?"	Excavate at $(0.48,\ 0.65)$
"Position the bucket for loading"	Position implement at detected pile $(0.62,\ 0.57)$
"Move to dumping area and unload"	Navigate to discharge zone $\rightarrow$ position for discharge $\rightarrow$ actuate discharge $\rightarrow$ reverse clear
"Go to the next pile"	Cautious approach to pile at $(0.71,\ 0.60)$
"What action should be taken?"	Navigate to highest-confidence detected pile

The multi-step command handler in during inference maps several composite instruction patterns (for example, pile-loading and dump-sequence prompts) to predefined ordered action programs. In our prototype, this deterministic layer is important because generated model outputs can vary in formatting and specificity, while downstream control interfaces require consistently parseable commands. The result is a practical hybrid design: learned semantic guidance from the language model, with explicit structure and validity guarantees from symbolic post-processing.

D. Fallback Mechanism Analysis

In this prototype, we treat fallback handling as a safety mechanism rather than a secondary feature. The parser hierarchy (pattern-based sequence handler, structured regex extraction, then keyword-based defaults) ensures that the system returns a well-formed action string even when raw model text is ambiguous. For early-stage systems operating on limited data, this explicit robustness layer is often more informative than reporting only end-model quality metrics.

V. Discussion

A. Implications for Long-Term Field Deployment

Reliable autonomous operation over extended durations in complex, unstructured environments is directly implicated by each component of our system, and the limitations we identify map cleanly onto core long-term autonomy research themes.

Distributional shift and environmental change. A wheel loader operating across a full working day encounters dramatic appearance variation: early-morning frost, midday sun, late-afternoon shadow, and post-rain mud. The present system processes each frame independently without modelling appearance change over time. Continual adaptation mechanisms [18] that update perception and action models as operational statistics evolve are a necessary extension for robust long-term deployment.

Persistent scene understanding. As material is loaded and deposited, pile positions and quantities change continuously. The current system builds no persistent world model, repeating the full detection procedure on each frame. Integration with a maintained environmental representation – even a lightweight occupancy map updated as piles are consumed – would support longer-horizon task planning and more efficient multi-cycle operation.

Safety, uncertainty, and interpretability. When the system processes a scene that falls outside its training distribution – an unusual material type, severe weather occlusion, a novel equipment configuration – it defaults silently to a rule-based fallback without communicating that its perception or action prediction is unreliable. Long-term field deployment demands explicit uncertainty quantification, enabling operators to provide clarifying guidance when the system's confidence is low [15], [19]. This is a precondition for justified operational trust in autonomous construction systems.

B. Limitations and Future Work

Training corpus scale and diversity. The three-sequence corpus used in this work represents only a narrow slice of the visual diversity encountered on real construction sites. Expanding the corpus to dozens or hundreds of sequences spanning multiple sites, machine types, material categories, and seasonal conditions is the most direct path to improved generalisation. Synthetic data augmentation from physics-based construction simulators could supplement real footage in a sim-to-real transfer framework [23].

Quantitative benchmarking. The present evaluation is intentionally qualitative, reflecting the system-description scope of this paper. The natural next step is the construction of a held-out, hand-labelled in-domain benchmark that supports (i) per-tier precision/recall and latency measurements for each stage of the perception cascade, and (ii) action-prediction accuracy against ground-truth operator action labels. Establishing such a benchmark will allow direct comparison with VLA baselines and quantitative ablation of the hybrid design choices presented here.

Metric spatial grounding. Target positions are currently expressed as normalised image-plane coordinates. Real robot deployment requires transformation to metric workspace coordinates, necessitating integration with depth sensing or monocular depth estimation [22] to recover three-dimensional pile positions from monocular imagery.

Closed-loop execution and reactivity. The current pipeline operates open-loop: it generates a single action for each input frame without observing the outcome of previous actions. Closed-loop operation – in which robot progress, unexpected obstacles, and pile depletion are continuously monitored and fed back into the action generation process – is essential for robust long-horizon task completion.

Ground-truth evaluation. Quantitative evaluation of action quality is fundamentally constrained in the absence of ground-truth action trajectories aligned with the source footage. Future work should collect paired datasets of construction site imagery with corresponding expert teleoperation demonstrations, enabling principled evaluation of action accuracy and spatial grounding quality.

VI. Conclusion

We have presented an initial hybrid vision-language prototype for natural-language-guided construction domain wheel-loader assistance in unstructured scenes. The implementation integrates data preparation, automated training-pair generation, parameter-efficient model adaptation, structured inference, and an interactive demo interface. This makes the project a interesting baseline for early research rather than a claim of deployable field autonomy. The study is based on our preliminary investigation and ongoing work in construction robotics, and the report should be interpreted as preliminary evidence. The current evidence suggests that hybrid design choices – especially perception cascades and deterministic action parsing – are useful for maintaining valid outputs under sparse data and noisy inputs. At the same time, the prototype remains limited by small corpus scale, image-plane (not metric) grounding, open-loop execution, and lack of calibrated uncertainty estimates. Addressing these gaps – together with the establishment of a labelled in-domain benchmark for per-tier perception metrics and action-prediction accuracy – is the next step toward rigorous evaluation on broader and more realistic datasets, followed by eventual real-machine validation.

REFERENCES

[1] N. Hawes *et al.*, "The STRANDS project: Long-term autonomy in everyday environments," in *IEEE Robotics & Automation Magazine*, 2017.

[2] B. Zitkovich *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Proceedings of The 7th Conference on Robot Learning (CoRL)*, pp. 2165–2183, 2023.

[3] D. Driess *et al.*, "PaLM-E: An embodied multimodal language model," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 8469–8488, 2023.

[4] M. J. Kim *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model." Conference on Robot Learning. PMLR, 2025.

[5] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li, "Learning to Act Anywhere with Task-centric Latent Actions," in *Proc. Robotics: Science and Systems (RSS)*, Los Angeles, CA, USA, June 2025.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, vol. 139, pp. 8748–8763, 2021.

[7] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, and N. Houlsby, "Simple open-vocabulary object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[8] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," `https://github.com/ultralytics/ultralytics`, 2023.

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 12888–12900, 2022.

[10] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 19730–19742, 2023.

[11] A. Grattafiori *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[13] M. Ahn *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," in *Proc. Conf. Robot Learning (CoRL)*, 2022.

[14] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2023.

[15] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[16] K. Doherty, T. Shan, J. Wang, B. Englot, "Learning-Aided 3-D Occupancy Mapping With Bayesian Generalized Kernel Inference," in *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 953–966, 2019.

[17] T. Krajník, J. P. Fentanes, J. M. Santos and T. Duckett, "FreMEn: Frequency Map Enhancement for Long-Term Mobile Robot Autonomy in Changing Environments," *IEEE Transactions on Robotics*, vol. 33, no. 4, pp. 964–977, 2017.

[18] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, 2019.

[19] N. Sünderhauf *et al.*, "The limits and potentials of deep learning for robotics," *The International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 405–420, 2018.

[20] A. Kirillov *et al.*, "Segment anything," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2023.

[21] W. Fang, L. Ding, H. Luo, P. E. D. Love, "Falls from heights: A computer vision-based approach for safety harness detection," in *Automation in Construction*, vol. 91, pp. 53–61, 2018.

[22] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 12159–12168, 2021.

[23] J. Tobin *et al.*, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2017.

[24] M. McCann, "Heavy equipment and truck-related deaths on excavation work sites," *Journal of Safety Research*, vol. 37, no. 5, pp. 511–517, 2006.