Grenze International Journal of Engineering and Technology, June Issue

RoboReward: Optimizing Vision-Language Reward Models for Scalable Robotic Learning

Ananya Manoj[1], Shamiul Mirza[2], Ujjawal Gupta[3], Rajveer S Yaduvanshi[4] and Manisha[5]
[1-5]Department of Instrumentation and Control Engineering, Netaji Subhas University of Technology, New Delhi, India
Email: ananya.manoj.ug22@nsut.ac.in, shamiul.mirza.ug22@nsut.ac.in, Ujjawal.gupta.ug22@nsut.ac.in, rajveer.yaduvanshi@nsut.ac.in, manisha.singh@nsut.ac.in

***Abstract*— Designing robust and scalable reward functions remains a significant challenge in reinforcement learning for robotic systems. Traditional reward engineering approaches are task-specific, labor-intensive, and fail to generalize across diverse real-world environments. In this work, we introduce RoboReward, a vision-language reward model that leverages pretrained CLIP embeddings to generate scalar reward signals from visual observations and natural language instructions.**

To improve computational efficiency, we propose an 8-frame temporal aggregation strategy that reduces redundancy while preserving essential task-relevant information. Compared to full-frame processing, this reduces training time from 6.48 days to approximately 2 hours on a single T4 GPU.

Experimental results demonstrate strong performance, achiev-ing a Pearson correlation of 0.896 with ground-truth rewards and a Mean Squared Error (MSE) of 0.2048. Furthermore, RoboReward generalizes across multiple tasks without requiring manual reward engineering.

These results highlight the effectiveness of vision-language models for scalable and efficient reward modeling in robotic learning systems.

***Index Terms*—Robotics, Reinforcement Learning, Vision-Language Models, Reward Modeling, CLIP.**

I. INTRODUCTION

Reinforcement learning (RL) has emerged as a powerful framework for enabling autonomous decision-making in robotic systems. It allows agents to learn optimal behaviors through continuous interaction with their environment. A fundamental component of RL is the design of reward functions, which provide the supervisory signal guiding policy optimization [1]. However, in practical robotic applications, reward functions are typically handcrafted and highly task-specific, requiring substantial domain expertise and iterative tuning. As tasks grow more complex and diverse, manual reward engineering becomes a significant bottleneck, limiting scalability and hindering deployment in dynamic, real-world environments.

This limitation is particularly evident in tasks involving high-level semantic objectives, long-horizon dependencies, and unstructured environments, where it is challenging to explicitly define reward functions. Prior approaches, such as inverse reinforcement learning and preference-based learning [2], [3], have attempted to address this issue. However, these methods often rely on large amounts of labeled data or human feedback, which restricts their practicality and scalability.

Grenze ID: 01.GIJET.12.2.263_26

Recent advances in Vision-Language Models (VLMs), such as CLIP [4], have demonstrated strong capabilities in aligning visual inputs with natural language semantics. These models have been extended to robotic applications, enabling instruction-following and generalization across tasks through multimodal reasoning [5]–[7]. By leveraging pretrained multimodal repre-sentations, VLMs provide a promising approach for learning reward functions directly from high-level task descriptions, reducing the dependence on manual reward engineering.

The core problem is the lack of scalable, generalizable reward functions for robotic RL. Manual reward engineering is impractical for complex tasks, while prior data-driven approaches (IRL, RLHF) require large expert datasets, limiting scalability. This motivates a reward model that requires no manual engineering, generalizes via natural language, and operates within limited compute. To address this, we introduce RoboReward, a vision-language reward model mapping video observations and natural language instructions to scalar rewards via pretrained CLIP embeddings, with an 8-frame temporal aggregation strategy that cuts training from multiple days to ~2 hours on a single T4 GPU.

Unlike prior approaches requiring large-scale fine-tuning or complex temporal architectures, RoboReward shows that simple frame averaging with frozen CLIP embeddings is sufficient for scalable reward modeling, offering a practical path to deployment on limited hardware.

Our main contributions are as follows:

- We propose RoboReward, a scalable vision-language reward modeling framework that eliminates the need for handcrafted reward functions.
- We introduce an efficient 8-frame temporal aggregation strategy, cutting training time from multiple days to under two hours without compromising performance.
- We demonstrate that RoboReward achieves competi-tive accuracy while significantly reducing training time, highlighting a favorable efficiency-performance trade-off, achieving a correlation of 0.896 with ground-truth rewards and a Mean Squared Error (MSE) of 0.2048.
- We validate the generalization capability of our approach across multiple tasks, highlighting its applicability to real-world robotic systems.

II. Literature Survey

A. Reward Learning in Reinforcement Learning

Reward modeling plays a critical role in reinforcement learning, particularly in settings where explicit reward design is difficult. Traditional approaches rely on manually engineered reward functions, which are often brittle and task-specific. To address this, prior work has explored learning rewards from demonstrations and human feedback. Inverse Reinforcement Learning (IRL) [2] aims to recover reward functions from expert behavior, while Reinforcement Learning from Human Feedback (RLHF) [3] learns reward models based on preference comparisons. Although these approaches reduce the need for manual reward design, they typically require large amounts of expert demonstrations or human annotations, limiting scalability in real-world applications.

B. Vision-Language Models for Representation Learning

Recent advances in Vision-Language Models (VLMs) have enabled joint reasoning over visual and textual modalities. A key breakthrough in this domain is Contrastive Language-Image Pretraining (CLIP) [4], which learns a shared embedding space by aligning images with their corresponding textual descriptions. This multimodal alignment allows models to generalize across tasks without task-specific supervision. CLIP and similar mod-els have demonstrated strong zero-shot performance across a wide range of vision tasks, making them a powerful foundation for downstream applications.

C. Multimodal Learning in Robotics

Building upon VLMs, recent research has explored their integration into robotic systems. Approaches such as SayCan

[5] combine language models with affordance-based reasoning to enable task planning, while PaLM-E [6] incorporates multimodal inputs directly into large-scale embodied mod-els. Similarly, VIMA [7] demonstrates the effectiveness of multimodal prompts for general robot manipulation tasks. These methods highlight the potential of combining vision and language to improve generalization and reduce task-specific engineering in robotics.

D. Limitations of Existing Approaches

Despite these advancements, several challenges remain. First, many existing approaches focus on policy learning rather than reward modeling, leaving the problem of scalable reward generation underexplored. Second, methods that leverage multimodal representations often require extensive fine-tuning or large-scale training, resulting in high computational costs.

Finally, temporal reasoning in video-based reward estimation is often handled inefficiently, leading to increased latency and resource consumption.

Furthermore, existing methods often prioritize model com-plexity over efficiency, leaving the design of lightweight and scalable reward models relatively underexplored.

E. Our Contribution

In contrast to prior work, we propose RoboReward, a lightweight and computationally efficient vision-language re-ward model that prioritizes scalability without sacrificing performance. By leveraging pretrained CLIP embeddings and a simple yet effective temporal aggregation strategy, our approach reduces computational overhead while maintaining strong alignment with task semantics.

III. Methodology

A. Problem Formulation

We aim to learn a vision-language reward function that maps visual observations and natural language instructions to scalar reward values. Given a trajectory $\tau = (v_1, v_2, ..., v_T)$ and a task description l, the objective is to learn:

$$R = f(\tau, l) \tag{1}$$

where $R \in \mathbb{R}$ represents how well the observed behavior aligns with the intended task.

In this work, we focus on supervised reward learning using datasets with explicit reward annotations. While the proposed framework can be extended to multi-dataset settings or alternative supervision signals, such extensions are beyond the scope of this study.

B. Dataset and Preprocessing

The dataset is based on publicly available robotic manipu-lation benchmarks, ensuring diversity in task objectives and visual contexts.

We utilize a dataset of approximately 10,000 video samples with corresponding task descriptions and reward annotations. From each trajectory, N = 8 frames are uniformly sampled to capture temporal context efficiently. Each frame is resized to
224 × 224 pixels.

C. Multimodal Feature Extraction

We use a pretrained CLIP model as a frozen backbone to extract visual and textual embeddings.
For each sampled frame:

$$z^v = \phi_v(v_i), \qquad z^v \in \mathbb{R}^{512} \tag{2}$$

For the language instruction:

$$z^l = \phi_l(l), z^l \in \mathbb{R}^{512} \tag{3}$$

D. Temporal Aggregation

Frame-level embeddings are aggregated using averaging:

$$\bar{z}^v = \frac{1}{N}\sum_{i=1}^{N} z_i{}^v \tag{4}$$

This provides a compact and robust representation of the input trajectory.

E. Feature Fusion

The aggregated visual embedding $\bar{z}^v$ and textual embedding z^l are fused into a joint representation. In our implementation, we use concatenation:

$$z = [\bar{z}^{v}; z^{l}] \quad (5)$$

While more advanced fusion mechanisms such as cross-attention may further improve alignment between modalities, we adopt concatenation for computational efficiency.

As illustrated in Fig. 1, the proposed pipeline integrates visual and language representations to predict reward signals.

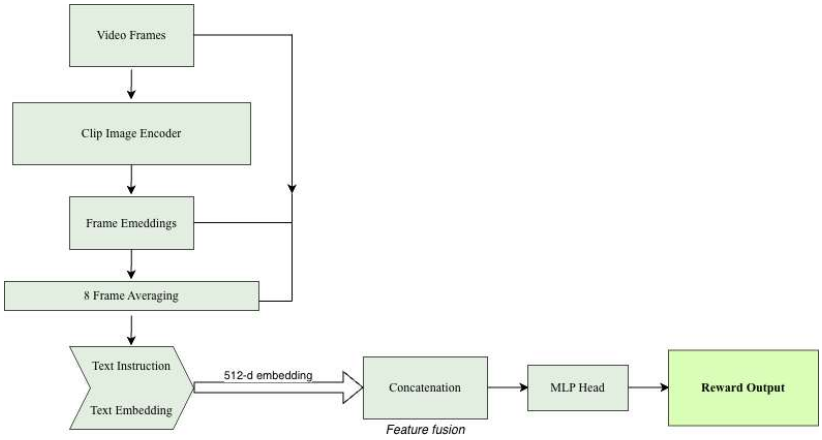

Fig. 1. Overview of the RoboReward framework

A sequence of video frames is processed using a CLIP-based visual encoder to extract frame-level embeddings, which are temporally aggregated. In parallel, task instructions are encoded using a CLIP text encoder. The resulting embeddings are fused and passed through a lightweight prediction network to generate scalar reward values.

F. Reward Prediction Network

The fused representation is passed through a lightweight Multi-Layer Perceptron (MLP):

$$R_{\text{pred}} = g(z) \quad (6)$$

where $g(\cdot)$ consists of:

$1024 \rightarrow 512 \rightarrow 1$

with ReLU activation. This design balances expressiveness and efficiency.

G. Reward Normalization

To stabilize training, reward values are normalized:

$$R_{\text{norm}} = \frac{R - \mu}{\sigma} \quad (7)$$

H. Training Objective

We primarily optimize the model using Mean Squared Error (MSE) loss:

$$L_{MSE} = \frac{1}{M} \sum_{i=1}^{M} (R_{Pred} - R_{Norm})^2 \quad (8)$$

Additionally, alternative training objectives such as classification-based or ranking-based losses can be incorporated to improve robustness across heterogeneous datasets. However, in this work, we focus on the regression objective for stability and simplicity.

I. RL Integration

The learned reward model can be integrated into reinforce-ment learning pipelines in future work.

$$R_{\text{total}} = \alpha r_{\text{env}} + \beta R_{\text{pred}} \quad (9)$$

where α and β control the contribution of environment and learned rewards. In this work, we focus on standalone reward modeling and leave full RL integration for future exploration

J. Computational Efficiency

RoboReward is designed for efficiency. By using pretrained CLIP embeddings, freezing the backbone, and reducing tem-poral redundancy through 8-frame sampling, training time is reduced from 6.48 days to approximately 2 hours on a single T4 GPU.

K. Implementation Details

All embeddings are L2-normalized prior to fusion. The CLIP backbone remains frozen during training. The model is trained using batch optimization with the following settings:

- Optimizer: Adam
- Learning rate: 5 × 10−5
- Batch size: 32
- Epochs: 25
- Input resolution: 224 × 224

IV. Extended Evaluation Across Benchmark Datasets

We evaluate the performance of RoboReward in terms of prediction accuracy, computational efficiency, and general-ization capability. Across multiple training runs, the model demonstrates consistent performance with negligible variance in both MSE and correlation metrics.

TABLE I: PERFORMANCE COMPARISON OF ROBOREWARD

Method	MSE ↓	Corr. ↑	Time ↓
Full Frame	0.198	0.902	6.48 days
CLIP Baseline	0.312	0.742	3 hours
RoboReward (Ours)	0.2048	0.896	2 hours

A. Ablation Study

Table II presents a quantitative comparison of different frame sampling strategies. Using fewer frames reduces computation but may lose important temporal information, while using full video sequences increases training cost significantly. Empirically, we evaluate frame sampling strategies using 4, 8, and full-frame inputs. While 4-frame sampling reduces computational cost further, it leads to a noticeable drop in correlation. Full-frame processing achieves slightly better accuracy but incurs significantly higher training time. The proposed 8-frame strategy provides the best trade-off between efficiency and performance.

TABLE II: ABLATION STUDY ON FRAME SAMPLING STRATEGIES

Number of Frames	MSE ↓	Correlation ↑	Training Time ↓
4 Frames	0.245	0.842	1.5 hours
8 Frames (Ours)	0.2048	0.896	2 hours
Full Frames	0.198	0.902	6.48 days

While full-frame processing achieves marginally better accu-racy, the improvement is minimal compared to the substantial increase in computational cost, highlighting the efficiency advantage of the proposed approach.

B. Computational Efficiency

A primary objective of RoboReward is to improve the computational efficiency of reward modeling. In the standard setup, handling full video sequences with around 640 frames each required about 6.48 days of training time on a T4 GPU. However, by using the proposed 8-frame temporal aggregation method, the training time dropped to roughly 2 hours. This demonstrates a significant improvement in computational efficiency.

These results indicate that temporal compression can signifi-cantly reduce computational requirements without compromis-ing essential information for reward prediction.

C. Training Stability

To analyze optimization behavior, we plot the training loss across epochs as shown in Fig. 2. The model demonstrates rapid convergence within the initial training phase, with loss decreasing sharply in the first few

epochs and stabilizing thereafter. This indicates that the combination of pretrained CLIP embeddings and reward normalization enables stable and efficient learning without gradient instability.

D. Quantitative Performance

By the 25th training cycle, the model reached a Mean Squared Error (MSE) of 0.2048 and a Pearson correlation coefficient of 0.896 between the predicted and actual reward values. The high correlation suggests that the model accurately understands how visual observations relate to the meaning of the task, while the low MSE indicates that the reward predictions are very accurate.

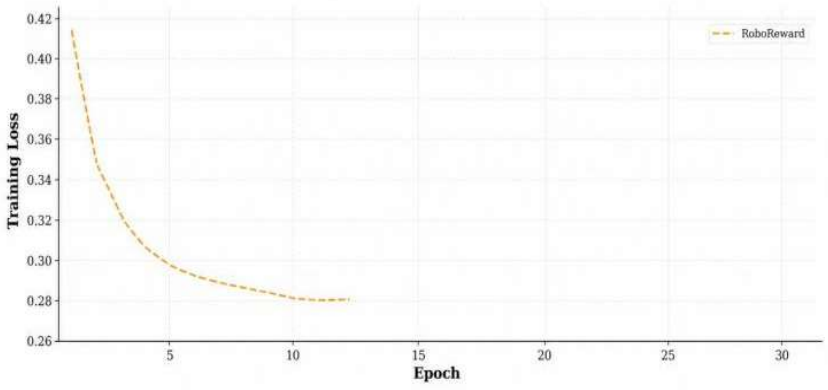

Fig. 2. Training loss curve of RoboReward across epochs. The model converges rapidly within the initial training phase and stabilizes without oscillations, indicating stable optimization behavior.

E. Qualitative Analysis

To further evaluate model performance, qualitative analysis was conducted on randomly sampled test cases.
For example, in one case, the model predicted a reward of 1.22 compared to the real value of 1.00, showing a close match.
Across multiple samples, predicted rewards consistently align with ground-truth values, indicating reliable generalization across diverse task scenarios. These results suggest that RoboReward demonstrates strong generalization to unseen samples, providing reliable reward estimates in various situations.

F. Experimental Setup

We evaluate the proposed RoboReward model on multiple robotic manipulation datasets, including LIBERO (Object, Spatial, Goal, and Long-Horizon variants), ALOHA, PushT, and the RoboReward dataset. These benchmarks cover a wide range of task complexities, including object manipulation, spatial reasoning, and long-horizon planning.
The datasets are obtained from publicly available sources on Hugging Face [8]–[14].

TABLE III: ROBOREWARD MODEL CONFIGURATION

Property	Value
Visual Backbone	CLIP ViT-B/32 (frozen)
Text Encoder	CLIP Text Transformer (frozen)
Embedding Dimension	512
Temporal Sampling	8 frames (uniform)
Fusion	Concatenation
Reward Head	MLP (1024 → 512 → 1)
Input Resolution	224 × 224
Loss Function	Mean Squared Error (MSE)
Optimizer	Adam
Batch Size	32
Epochs	25
Hardware	NVIDIA T4 GPU

Each dataset is divided into training and evaluation sets using an 80:20 split. All reported results are computed on held-out evaluation data to ensure unbiased performance assessment.

TABLE IV: PERFORMANCE OF ROBOREWARD ACROSS MULTIPLE ROBOTIC BENCHMARK DATASETS

Dataset	AUROC ↑	RMSE ↓
LIBERO-Object	0.9406	0.2004
LIBERO-Spatial	0.8907	0.2391
LIBERO-Goal	0.7611	0.2905
LIBERO-Long	0.9270	0.2087
ALOHA	0.9400	0.1047
PushT	0.5032	0.2922
RoboReward	0.7190	0.4805

G. Benchmark Datasets

The results demonstrate that RoboReward achieves strong performance on object-centric and long-horizon tasks, particu-larly on the LIBERO-Object and ALOHA datasets, achieving high AUROC scores with relatively low RMSE values. How-ever, performance degrades significantly on PushT, where the model operates close to random behavior, indicating limitations in capturing fine-grained geometric reasoning tasks.
Additionally, moderate performance on the RoboReward dataset suggests challenges in handling diverse reward scales and multi-robot environments.

H. Failure Analysis

The model exhibits limitations in tasks requiring precise geometric reasoning. In such cases, performance degrades significantly, suggesting that CLIP-based representations may not capture fine-grained spatial relationships effectively.
Additionally, tasks involving diverse multi-robot scenarios show reduced accuracy, indicating that the current model struggles with heterogeneous environments and complex reward distributions.
These observations highlight important limitations of vision-language reward models and suggest directions for future improvements.

I. Limitations

The current evaluation is based on a limited number of runs, and additional experiments with multiple random seeds are required to assess robustness. Furthermore, the reliance on simple temporal averaging may limit performance on tasks requiring detailed temporal reasoning. Finally, the absence of diverse baseline comparisons restricts a comprehensive understanding of the model's relative performance.

J. Discussion

The results demonstrate two key advantages of the proposed approach. First, RoboReward achieves a strong balance between computational efficiency and prediction accuracy, making it suitable for scalable robotic learning.
The 8-frame temporal aggregation strategy further demon-strates robustness across varying task durations, indicating that uniform sampling is sufficient to capture essential task progression.
Overall, these findings confirm that RoboReward provides a practical and scalable alternative to traditional reward engineering in reinforcement learning.
While direct comparison with large-scale vision-language reward models is limited due to differences in evaluation protocols and computational requirements, our results demon-strate competitive performance relative to lightweight baselines, emphasizing the efficiency-performance trade-off achieved by RoboReward.

V. CONCLUSION AND FUTURE WORK

In this work, we introduced RoboReward, a scalable vision-language reward modeling framework that addresses the limitations of manual reward engineering in reinforcement learning.
Our experiments demonstrate strong performance, achieving a Pearson correlation of 0.896 and a Mean Squared Error (MSE) of 0.2048, while reducing training time from several days to approximately two hours on a single T4 GPU.
One limitation of the proposed approach is the reliance on simple temporal averaging, which may not capture complex temporal dependencies in long-horizon tasks.
Future work will focus on expanding the dataset to improve robustness, exploring attention-based temporal modeling tech-niques, and integrating RoboReward into full reinforcement learning pipelines for real-world robotic deployment like au-tonomous drones. UAVS (Unmanned Aerial Vehicles), Robotic arm manipulators, etc.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Netaji Subhas University of Technology (NSUT) for providing the necessary resources and support for this research.

REFERENCES

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. MIT Press, 2018.
[2] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in Proceedings of the International Conference on Machine Learning (ICML), 2000, pp. 663–670.
[3] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh et al., "Learning transferable visual models from natural language supervision," in Proceedings of the International Conference on Machine Learning (ICML), 2021.
[5] M. Ahn, A. Brohan et al., "Do as i can, not as i say: Grounding language in robotic affordances," arXiv preprint arXiv:2204.01691, 2022.
[6] D. Driess, F. Xia et al., "Palm-e: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.
[7] Y. Jiang et al., "Vima: General robot manipulation with multimodal prompts," in Proceedings of the International Conference on Learning Representations (ICLR), 2023.
[8] "Libero object dataset," https://huggingface.co/datasets/lerobot/libero_ object_image.
[9] "Libero spatial dataset," https://huggingface.co/datasets/lerobot/libero_ spatial_image.
[10] "Libero goal dataset," https://huggingface.co/datasets/lerobot/libero_goal image.
[11] "Libero long dataset," https://huggingface.co/datasets/lerobot/libero_10 image.