A Framework for Accurate and Efficient Image Captioning by Fusing Fine-Tuned YOLOv8 and LLMs

Dabbrata Das [1*], Md Rishadul Bayesh [2], Mohammad Amanul Islam [3], Md. Asifur Rahman [4]

[1]Uttara University
Holding 77, Beribadh Road, Turag, Uttara, Dhaka 1230, Bangladesh

[1]dasdabbrata@gmail.com, [2]rishadul.b@uttarauniversity.edu.bd

[3]amanul.islam@uttarauniversity.edu.bd, [4]asifur.rahman@uttara.ac.bd

Abstract. *Automatic image captioning is essential for generating natural language descriptions by extracting meaningful features and understanding contextual relationships in images. While traditional methods like CNN-RNN models struggle with computational complexity and spatial awareness, multimodal Large Language Models (LLMs) offer an alternative but often lack object precision and are computationally expensive. In this work, we propose a novel image captioning framework that combines a fine-tuned YOLOv8 model with an LLM for efficient and accurate caption generation. YOLOv8 detects objects, extracts their names, confidence scores, and bounding box coordinates, which are filtered based on confidence levels above 0.5 before being passed to the LLM. This integration results in richer, more contextually accurate captions with lower inference time compared to existing methods. We evaluate our approach against multimodal LLMs and CNN-RNN models, demonstrating that it significantly improves efficiency while maintaining high caption quality. This method provides a promising solution for real-time applications, offering faster and more reliable image captioning for systems such as autonomous technologies and content generation.*

1 Introduction

Automatic image captioning is a fundamental task in computer vision that aims to generate natural language descriptions of images by extracting meaningful features and understanding the contextual relationships within a scene. Traditional approaches, such as Convolutional Neural Networks (CNNs) combined with Recurrent Neural Networks (RNNs) utilizing Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU), have been widely adopted for this task. More recently, multimodal Large Language Models (LLMs) have emerged, demonstrating the capability to generate captions directly from images. However, these models often suffer from limitations such as the inability to maintain precise object details and spatial awareness, while also being computationally expensive with longer inference times, making them less efficient for real-time applications.

In this study, we propose a novel image captioning framework that integrates a fine-tuned YOLOv8 [1] model with an LLM to generate contextually accurate, detailed, and efficient image descriptions. We utilize the ADE20K [2] dataset to fine-tune a pre-trained YOLOv8 [1] model for object detection, enabling precise identification of objects along with their confidence levels and bounding box coordinates. To ensure reliability, we filter detected objects based on a confidence threshold of 0.5, passing only high-confidence detections to the LLM [3] for caption generation. By incorporating spatial information alongside object labels, positions, and confidence levels, our method enables the LLM [3] to produce more structured and descriptive captions that accurately reflect the scene.

To evaluate the effectiveness of our proposed approach, we conduct a comparative analysis against two widely recognized captioning methods: (1) multimodal LLM-based caption generation, which directly generates captions from images, and (2) CNN-RNN-based [4] models, where feature extraction via CNNs is followed by sequence generation using LSTMs or GRUs. Through extensive experiments, we demonstrate that our fine-tuned YOLOv8 + LLM framework generates captions that not only preserve the semantic integrity of objects but also maintain a standard level of captioning quality. Our results highlight the advantages of incorporating structured object detection data into LLM-based captioning, leading to improved accuracy and informativeness compared to existing approaches. The key contributions of our work include:

- The integration of a fine-tuned YOLOv8 model with an LLM to enhance image captioning performance.
- A confidence-based filtering mechanism to ensure accurate and reliable object detection before caption generation.
- A structured approach that incorporates object labels, confidence scores, and spatial positions to improve caption informativeness.
- A comparative evaluation demonstrating the efficiency of our method against multimodal LLM-based and CNN-RNN-based captioning models.

Our approach aligns with ongoing advances in data-driven artificial intelligence, emphasizing the importance of structured visual input for generating high-quality image descriptions. We believe that this framework can serve as a foundation for future research in vision-language modeling and real-world applications such as assistive technologies and content generation.

*Corresponding author
Data: https://www.kaggle.com/datasets/awsaf49/ade20k-dataset

2 Related Works

Recent approaches in image captioning leverage multimodal architectures to enhance semantic richness, though they differ significantly in spatial precision, computational efficiency, and real-time applicability. Rotstein et al. [5] introduced FuseCap, combining LLMs with object detection and OCR to enrich captions, yet it struggled with capturing spatial relationships. Yang et al. [6], through LDRE, improved visual reasoning by ensemble captioning, but at the expense of heavy computational overhead, unlike the simpler but spatially limited approach of FuseCap. Kim et al. [7] presented MLBCAP, using multiple LLMs for scientific figure captioning, emphasizing informativeness and clarity. However, MLBCAP relied on closed-source models and faced computational limitations similar to those in Patel et al. [8], whose Meta Llama-3.2-11B-Vision also exhibited latency issues despite broad domain applicability. Chen et al. [9] developed Attention Buckets, enhancing LLM attention mechanisms internally, thus improving contextual understanding. This internal modification differed from input-based methods like FuseCap or LDRE but incurred greater computational complexity and memory requirements. Li et al. [10] proposed CAT-LLM, optimized for retrieval tasks but compromised inference speed and lacked spatial object structuring. Matsuyama et al. [11]'s LaVCa addressed fMRI-based captioning with high interpretability but did not handle visual scenes or spatial context. Other approaches, like Admoni et al.'s [12] SyS-LLM, Bhambri et al.'s [13] MEDIC, and Sun et al.'s [14] MDP3, focused on specialized domains such as reinforcement learning or video frame selection, diverging from real-time captioning of static visual scenes. In contrast, our proposed method uniquely integrates efficient YOLOv8-based object detection with structured, spatially-aware inputs to an LLM, addressing prior gaps in spatial precision, inference efficiency, and suitability for real-time applications. Table 1 shows the comparison of the recent works with the proposed method.

3 Methodology

This section provides a detailed overview of the network architecture, and the complete workflow.

3.1 Dataset

In this method, we utilize two primary datasets to enhance object detection capabilities for our image captioning model. The first is the COCO (Common Objects in Context) [15] dataset, which serves as the foundation for the pre-trained YOLOv8 model. COCO [15] contains a diverse set of object categories and rich annotations, making it suitable for initial object detection tasks. To further enhance the model's ability to detect a wider variety of objects, particularly those not well-represented in COCO [15], we introduce the ADE20K [2] dataset. ADE20K [2] includes additional categories such as trees, rivers, mountains, sky, and other natural elements, which are crucial for generating more comprehensive and contextually accurate captions in outdoor and nature scenes. By fine-tuning the YOLOv8 [1] model with ADE20K [2], we enable the model to detect a broader range of objects, ensuring it performs effectively across diverse image scenarios. The combination of these datasets allows us to enrich object recognition and improve the overall captioning process, particularly for scenes involving natural landscapes and varied object types.

3.2 Model Architecture

To describe the network architecture (Figure 1) of the proposed model, we divide the process into three phases. First, we perform object detection using a fine-tuned YOLOv8 model to accurately identify objects in the image. Next, we extract relevant features along with key information such as object labels, confidence scores, and spatial positions. Finally, this extracted information is formatted as a structured prompt and passed to an LLM, enabling it to generate detailed and contextually accurate captions efficiently.

3.2.1 Object Detection Using Fine-Tuned YOLOv8

In the first phase of our framework, we utilize a fine-tuned YOLOv8 model for object detection. YOLOv8 [1] is chosen for its efficiency and accuracy in real-time object detection tasks. We fine-tune the model on the ADE20K [2] dataset to enhance its ability to detect a wide variety of objects with high precision. The model outputs detected objects along with their bounding box coordinates and confidence scores. To ensure only relevant and reliable objects are used for caption generation, we employ a confidence-based filtering mechanism, discarding detections below a predefined threshold (0.5). This step ensures that only meaningful objects contribute to the final caption, improving overall caption accuracy and informativeness.

3.2.2 Feature Extraction and Representation

Once objects are detected, we extract essential features, including object labels, confidence scores, and spatial positions. To ensure the reliability of detected objects, we apply a filtering step, retaining only those with a confidence score greater than 0.5. The filtered objects are then structured into a compact and meaningful representation, which serves as input for the captioning model. The spatial positions help the model understand object relationships within the image, leading to more informative and contextually accurate descriptions. The extracted features are formatted into a structured prompt to ensure compatibility with the LLM used for caption generation.

3.2.3 Caption Generation Using LLMs

In the final phase, the extracted object information is fed into a Large Language Model (LLM) to generate descriptive and coherent captions. The LLM [3] takes the structured prompt containing detected objects, their spatial relationships, and confidence scores as input. Using its natural language generation capabilities, the LLM produces fluent and contextually relevant captions that describe the image comprehensively. To further improve caption coherence and factual accuracy, we apply a prompt refinement

Table 1: Comparison of Related Works with the Proposed Method

Work	Approach	Limitations	Comparison to Our Work
Rotstein et al [5]	Combines LLM with object detection, attribute recognition, and OCR	Weak spatial relationship modeling between objects	Includes bounding box and spatial data to improve spatial awareness
Yang et al. [6]	Zero-shot image retrieval using divergent reasoning and caption ensemble	Computationally heavy due to multiple captions	Avoids ensemble generation; more efficient for real-time applications
Kim et al. [7]	Multi-LLM system for scientific figure captioning	Relies on closed-source LLMs; high computational cost	Uses a single LLM with efficient YOLOv8 input; more lightweight and accessible
Chen et al. [9]	Adjusts Rotary Position Embedding angles for better attention	Increases memory usage due to parallel inference streams	Enhances input structure instead of changing model architecture
Matsuyama et al. [11]	LLM-based voxel-level captioning from fMRI data	Domain-specific; lacks visual feature detail; resource-intensive	Optimized for general image scenes; designed for real-time use
Patel et al. [8]	Multimodal LLM for cross-domain captioning	High latency; unsuitable for real-time	Reduces latency by offloading detection to YOLOv8; enables faster captioning
Admoni et al. [12]	Uses LLMs for summarizing RL policies in natural language	Requires precise inputs; computationally demanding	Simplifies the pipeline; avoids RL complexity for general-purpose images
Bhambri et al. [13]	Integrates LLM with feedback critic in RL	High cost due to iterative feedback	Direct, single-pass caption generation without iterative loops
Sun et al. [14]	Frame selection with DPP and MDP for video captioning	Computationally intensive for long videos	Supports single-image captioning without temporal reasoning

strategy where the model generates multiple candidate captions, which are then summarized or re-ranked based on predefined quality metrics. This process ensures that the generated captions remain concise, informative, and well-aligned with the visual content of the image.

3.3 Evaluation Metrics

To assess the performance of our proposed model, we employ the following evaluation metrics:

- **BLEU-4 [16] and METEOR [17]:** Standard captioning metrics used to evaluate linguistic quality, coherence, and relevance of the generated captions.
- **CLIPScore [18] (Cosine Similarity):** A metric that measures the semantic alignment between the generated caption and the image content. It is computed using the cosine similarity between the image and caption embeddings in a shared space, as follows:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

where $\mathbf{A}$ is the vector representing the image, $\mathbf{B}$ is the vector representing the caption, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (norms) of the respective vectors.

Algorithm 1 Image Captioning Using Fine-Tuned YOLOv8 and LLM

Input: Image I
Output: Caption C
Step 1: Object Detection using YOLOv8
Detect objects in image I using fine-tuned YOLOv8.
Extract object labels, confidence scores, and bounding box coordinates.
Step 2: Caption Generation using LLM
Pass the retained object labels, confidence scores, and spatial positions to a pre-trained LLM.
Generate a descriptive caption based on the provided information.
Step 3: Confidence-Based Filtering
for each detected object **do**
  if confidence score >0.5 **then**
    Retain the object for further processing.
  else
    Discard the object.
  end if
end for
Step 4: Feature Extraction and Representation
Structure the extracted features into a prompt format for LLM input.
Step 5: Output the Caption
Return the generated caption C.

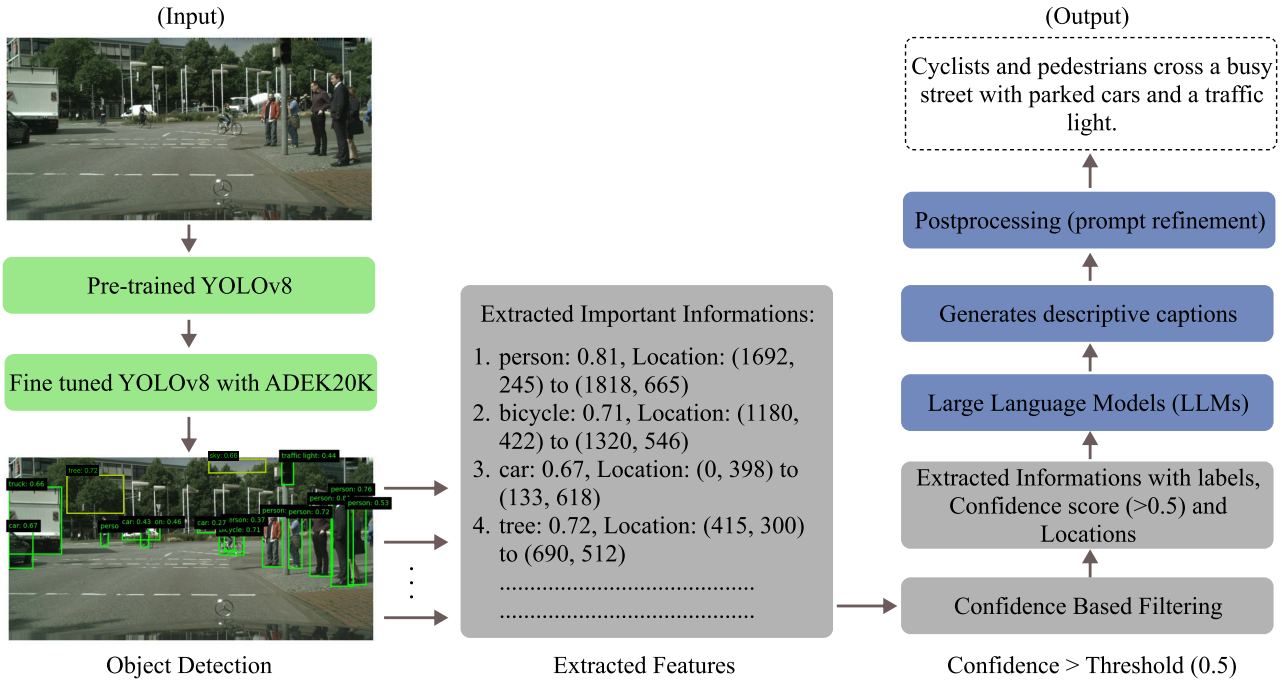

Figure 1: Architecture of the proposed image captioning model using fine-tuned YOLOv8 [1] and LLM [3] for efficient and contextually rich captions.

3.4 Experimental Setup

Our experiments were conducted on a Kaggle environment equipped with dual NVIDIA T4 GPUs. We fine-tuned the YOLOv8 [1] model using the ADE20K [2] dataset to enhance object detection capabilities. The model was trained for 5 epochs, leveraging CUDA acceleration for efficient processing. For caption generation, we employed the **microsoft/Phi-3-mini-4k-instruct** [3] large language model, utilizing the Hugging Face Transformers library. The structured object information, including labels, confidence scores, and spatial positions, was formatted as prompts and passed to the LLM for caption generation.

4 Result and Discussion

4.1 Quantitative Analysis

In this study, we evaluated the performance of machine-generated captions using three key evaluation metrics: **BLEU-4** [16], **METEOR** [17], and **CLIPScore** [18]. The results varied across different images and captioning models. In table 2, the proposed YOLOv8 + LLM model achieved the highest **BLEU-4** [16] score (0.166) for Image 1, indicating a stronger n-gram overlap with the reference captions compared to other models. However, for Image 2 and Image 3, the highest BLEU-4 scores were 0.226 and 0.548, respectively, achieved by different models. This suggests that while the proposed approach performs well in certain cases, traditional methods may still excel in others.

The **METEOR** [17] score, which accounts for stemming, synonym matching, and exact word overlap, showed the strongest performance for the proposed model in Image 1 (0.747). This highlights its ability to generate semantically similar captions. However, for Image 2, the CNN + RNN model outperformed others with a METEOR score of 0.457, while for Image 3, the proposed model achieved a score of 0.181, slightly improving over other approaches. The **CLIPScore** [18], which measures the alignment between the generated caption and the corresponding image in a multimodal embedding space, was consistently high across all models. The proposed YOLOv8 + LLM model achieved the highest CLIPScore (0.906) for Image 1, indicating superior semantic consistency with human perception. However, for Image 2 and Image 3, different models exhibited competitive performance, with CLIPScores reaching 0.887 and 0.854, respectively.

In summary, the proposed **YOLOv8 + LLM** model demonstrates strong performance, particularly in terms of METEOR [17] and CLIPScore [18], suggesting its ability to generate captions with high semantic similarity to human descriptions. While BLEU-4 [16] performance varies across images, the results indicate that different models may excel in different contexts.

4.2 Qualitative Analysis

In addition to the quantitative evaluation, a qualitative analysis was conducted to assess the clarity, relevance, accuracy, and overall quality of the generated captions. The model demonstrates a strong ability to capture the core content of images; however, it shows limitations in handling complex scenes, particularly in terms of grammatical correctness and detail richness. These qualitative observations are consistent with the quantitative findings, suggesting that while the captions are generally coherent and relevant, further improvements are needed to enhance accuracy and fluency. Notably, the model achieves efficient

Input RGB Image	Human Generated Caption	Machine Generated Caption		
		CNN+RNN	Multimodal LLM	YOLOv8 + LLM **(Proposed)**
1. 	Several different cars are parked along the side of the road.	Two cars are parked on a street.	A street with cars parked on it.	A group of cars is parked in different locations.
2. 	A young boy taking a bite of pizza.	A young boy in a black shirt is eating food.	A boy eating a piece of food.	A person enjoying a slice of pizza.
3. 	A boy watching boats with people on the water in the distance.	A young boy in a green shirt watches two boats.	A boy looking at boats.	A person is seen near a boat, while another person is further away.

Figure 2: Comparison of machine-generated and human-generated captions. The figure displays three images with their corresponding captions.

Table 2: Performance Comparison of Caption Generation Methods

Method	**Params**	**Inference Time (s)**	**BLEU-4** [16]	**METEOR** [17]	**CLIPScore** [18]
Image 1 (From figure 2)					
CNN + RNN [4]	4.32M	0.112	0.126	0.374	0.892
Multimodal LLM [19]	2.85B	3.09	0.125	0.241	0.858
YOLOv8 + LLM **(Proposed)**	3.80B	0.343	**0.166**	**0.747**	**0.906**
Image 2 (From figure 2)					
CNN + RNN [4]	4.32M	0.09	**0.226**	**0.457**	0.814
Multimodal LLM [19]	2.85B	1.512	0.193	0.253	0.885
YOLOv8 + LLM **(Proposed)**	3.80B	0.331	0.192	0.399	**0.887**
Image 3 (From figure 2)					
CNN + RNN [4]	4.32M	0.141	0.182	0.170	0.795
Multimodal LLM [19]	2.85B	3.35	**0.548**	0.166	**0.854**
YOLOv8 + LLM **(Proposed)**	3.80B	0.269	0.222	**0.181**	0.792

caption generation with faster inference time and, in certain cases, demonstrates superior object detection capabilities compared to other models. These insights underscore the modelâs potential and highlight areas for refinement to produce more precise, human-like, and fluent captions.

5 Future Works

In future work, we plan to further refine the YOLOv8 + LLM model to enhance its performance in generating even more contextually accurate captions. This can be achieved by experimenting with different multimodal architectures, including the integration of advanced transformers or self-attention mechanisms to improve context comprehension. Additionally, exploring domain-specific fine-tuning could help in adapting the model to specialized image datasets, thereby increasing its robustness and versatility. Future efforts will also focus on reducing inference time while maintaining or improving the quality of generated captions. Lastly, we aim to incorporate user feedback, especially from blind or visually impaired users, to ensure that the captions are truly effective in conveying essential image information in real-world applications.

6 Conclusion

In this study, we proposed the YOLOv8 + LLM framework for image caption generation, which balances accuracy and efficiency. While the CNN + RNN model [4] outperforms the proposed framework in terms of inference time, the YOLOv8 + LLM framework excels in generating contextually accurate and semantically aligned captions. It produces superior results across multiple evaluation metrics while maintaining competitive inference time, making it a well-rounded solution for real-world applications. The framework shows notable improvements in METEOR [17]

and CLIPScore [18], indicating strong alignment with human understanding, despite a slightly higher inference time compared to traditional methods. However, it faces limitations, such as the trade-off between speed and performance on certain images, where CNN + RNN remains more efficient for real-time applications. Further fine-tuning for specialized datasets could enhance its robustness and generalization capabilities.

In conclusion, the proposed framework offers a compelling balance of efficiency and accuracy, making it an excellent choice for applications that require high-quality captions while maintaining operational efficiency. Despite some limitations, the framework's ability to generate contextually rich and semantically accurate captions positions it as a valuable tool for a wide range of image captioning tasks.

Acknowledgment

This is the author's accepted manuscript of the paper published in the *Proceedings of the 2025 IEEE International Conference on Quantum, Photonics, Artificial Intelligence, and Nanotechnology (QPAIN)*. The final version is available at IEEE Xplore via DOI: https://doi.org/10.1109/QPAIN66474.2025.11171749

References

[1] R. Varghese and S. M., "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024, pp. 1–6.

[2] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," 2018. [Online]. Available: https://arxiv.org/abs/1608.05442

[3] M. Abdin *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024. [Online]. Available: https://arxiv.org/pdf/2404.14219

[4] S. Rohitharun, L. Uday Kumar Reddy, and S. Sujana, "Image captioning using cnn and rnn," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, 2022, pp. 1–8.

[5] N. Rotstein, D. Bensaid, S. Brody, R. Ganz, and R. Kimmel, "Fusecap: Leveraging large language models for enriched fused image captions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5689–5700.

[6] K. Wang, S. Lou, J. Wang, and X. Yuan, "Llm based semantic fusion of infrared and visible image caption," in *2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, vol. 9. IEEE, 2024, pp. 315–319.

[7] J. Kim, J. Lee, H.-J. Choi, T.-Y. Hsu, C.-Y. Huang, S. Kim, R. Rossi, T. Yu, C. L. Giles, T.-H. K. Huang *et al.*, "Multi-llm collaborative caption generation in scientific documents," *arXiv preprint arXiv:2501.02552*, 2025.

[8] V. Patel, A. Modi, H. Mistry, A. Mishra, R. Upadhyay, and A. Shah, "From alt-text to real context: Revolutionizing image captioning using the potential of llm," *arXiv preprint arXiv:2503.xxxxx*.

[9] Y. Chen, A. Lv, T.-E. Lin, C. Chen, Y. Wu, F. Huang, Y. Li, and R. Yan, "Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool use," *arXiv preprint arXiv:2312.04455*, 2023.

[10] W. Li, H. Fan, Y. Wong, M. Kankanhalli, and Y. Yang, "Cat-llm: Context-aware training enhanced large language models for multi-modal contextual image retrieval," *arXiv preprint arXiv:2403.xxxxx*.

[11] T. Matsuyama, S. Nishimoto, and Y. Takagi, "Lavca: Llm-assisted visual cortex captioning," *arXiv preprint arXiv:2502.13606*, 2025.

[12] S. Admoni, O. Ben-Porat, and O. Amir, "Sysllm: Generating synthesized policy summaries for reinforcement learning agents using large language models," *arXiv preprint arXiv:2503.10509*, 2025.

[13] S. Bhambri, A. Bhattacharjee, S. Kambhampati *et al.*, "Efficient reinforcement learning via large language model-based search," in *NeurIPS 2024 Workshop on Open-World Agents*, 2024.

[14] H. Sun, S. Lu, H. Wang, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and M. Li, "Mdp3: A training-free approach for list-wise frame selection in video-llms," *arXiv preprint arXiv:2501.02885*, 2025.

[15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. DollÃ¡r, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311â318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[17] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909/

[18] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," 2022. [Online]. Available: https://arxiv.org/abs/2104.08718

[19] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023. [Online]. Available: https://arxiv.org/abs/2301.12597