CS 6180 – Project Proposal
Semantic Diff Prompting for Video Understanding

Kaustubha Eluri, Jiameng Ji, Yifei Li, Siyi Gao

1 Introduction

Vision-Language Models (VLMs) such as GPT-4o, LLaVA, and Qwen-VL have improved video description quality, yet they still struggle with temporal consistency and redundant descriptions. Traditional captioning methods describe each frame independently or produce a single caption that ignores frame-level transitions. This leads to repeated content, hallucinated objects, and poor temporal grounding.

This project explores a simple but novel prompting strategy called **Semantic Diff Prompting**, where the model describes only the *semantic differences* between consecutive frames. Rather than re-describing static elements, the model focuses on meaningful changes, similar to how humans perceive motion.

We investigate whether Semantic Diff Prompting can (1) reduce hallucination, (2) improve temporal coherence, (3) reduce token redundancy, and (4) preserve or improve descriptive accuracy. Our work contributes to multimodal prompting, temporal reasoning, and efficient video understanding with potential applications in assistive technologies, video summarization, and resource-efficient inference.

2 Previous Work and Related Literature

2.1 Video Captioning and Temporal Modeling

Models such as VideoBERT and Masked Transformer for Video learn temporal embeddings to summarize motion. While effective for global reasoning, they do not focus on fine-grained frame-to-frame changes. Diff prompting directly targets these semantic transitions through prompt design rather than architectural changes.

2.2 Hallucination Reduction in VLMs

Studies show that VLMs hallucinate objects in cluttered or ambiguous scenes. Existing mitigation methods rely on external grounding modules, object detectors, or supervised filtering. Our approach aims to reduce hallucination through prompting alone, offering a lightweight, training-free alternative.

2.3 Structured Prompting and Lexical Conditioning

Prompt engineering techniques such as constrained decoding, dynamic lexicon injection, and chain-of-thought reasoning demonstrate that structured prompts shape model output significantly. Semantic Diff

Prompting extends this idea by imposing a temporal-difference constraint that encourages models to report only new or changed information.

3 Proposed Contributions

Our project focuses on three key contributions:

1. **Semantic Diff Prompting Framework**: We formalize a change-focused prompting strategy for video understanding.
2. **Empirical Evaluation**: We measure hallucination, temporal consistency, and token efficiency under diff prompting across several VLMs.
3. **Semantic Drift Score**: We introduce a metric to quantify deviation between predicted and ground-truth frame-level changes.

These contributions are feasible without model training and align with the time constraints of the course.

4 Methodology

4.1 Prompting Framework

We compare two prompting conditions:

- **Baseline**: "Describe this frame."
- **Semantic Diff Prompting**: "Describe only what changed since the previous frame. Do not repeat unchanged information."

We will curate 5–10 short videos (5–20 frames) representing both stable and dynamic scenes.

4.2 Evaluation Metrics

We evaluate:

- **Hallucination Rate**: Incorrect or nonexistent objects/actions.
- **Temporal Consistency**: Accuracy of change tracking across frames.
- **Token Efficiency**: Average tokens per frame.
- **Semantic Drift Score**: Our proposed change-based accuracy metric.

Evaluation uses manual annotation and LLM-based scoring.

4.3 Alternative Approaches Considered

We also considered:

- **Dynamic Lexicon Prompting**: Scene-specific keyword injection.
- **Chain-of-Thought for Motion**: Stepwise reasoning about transitions.
- **Scene Graph Diffing**: Object-relation graph comparison.

These methods may offer precision but require heavier engineering. Diff prompting remains the most lightweight and interpretable.

5 Expected Impact

Semantic Diff Prompting offers a minimal, training-free mechanism for improving temporal grounding, reducing redundancy, and mitigating hallucinations. The approach could support real-time systems, accessible video descriptions, and efficient multimodal inference. More broadly, it introduces a direction for temporal-aware prompting in generative video models.

References

- Sun, Chen, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. "VideoBERT: A Joint Model for Video and Language Representation Learning." ICCV, 2019. https://arxiv.org/abs/1904.01766
- Rohrbach, Anna, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. "Object Hallucination in Image Captioning." EMNLP, 2018. https://aclanthology.org/D18-1437/
- Mohanty, Anshuman et al. "The Future of MLLM Prompting is Adaptive: A Comprehensive Experimental Evaluation of Prompt Engineering Methods for Robust Multimodal Performance." Transactions on Machine Learning Research (TMLR). https://openreview.net/forum?id=B1L8HrjoA1