Paper Review - Zelda: Video Analytics using Vision-Language Models

Summary

This paper talks about Zelda, a video search system that uses a single zero-shot Vision Language Model (VLM) (like CLIP by OpenAI) to let users query large video data in natural language. It fixes common VLM issues such as prompt sensitivity (small wording changes, alter end results) and repetitive or low-quality results, by auto-expanding prompts (by adding synonyms, discriminator labels, and image quality cues (like blurry/grainy)) and by selecting top-K with diversity instead of plain cosine ranking. The system returns frames that are both relevant and semantically varied without per-query training. On five datasets (dashcam footage, camera footage, movie footage, news interviews and action footage) and 19 queries, Zelda improves diversity and accuracy over out-of-the-box VLM baselines and is substantially faster than a strong video analytics engine (VIVA) at matched quality.

Evaluation

With massive video data and recent ML advances, analysts need fast, flexible search without per-query model tweaks. The problem is significant, and the proposed approach is both valid and novel: keep a zero-shot VLM for simplicity, then add lightweight prompt expansion plus near duplicate removal and diversity aware selection to clean up the top-K using Maximal Marginal Relevance (MMR) (trades off relevance vs similarity to already selected frames). The empirical study is solid, with 5 datasets and 19 queries, clear baselines (CLIP-Relevant, CLIP-Diverse, VIVA), appropriate metrics (Mean Average precision(relevance), Average Pairwise Similarity(diversity)), and ablations that show each component's value. The writing and figures are clear, with a logical flow and clear explanations of the implementation, evaluation methodology, and metrics.

Main takeaways

- Zero-shot VLM search is fast because there is no per-query training, and model management is simpler since a single general model serves all queries.
- Out-of-the-box VLMs need help: quality and near-duplicate filtering with diversity selection.
- Zelda's approach: expand prompts to boost relevance, filter low-quality and near-duplicate frames, and select a diverse, high-confidence top-K using MMR

Strengths

- Zelda handles key limitations of video analytics (expressivity, model per predicate, redundancy, diversity) by using a zero-shot pipeline with auto-expanding prompts and diversity filtering.
- Zelda uses a semantic diversity pruning method based on VLM embeddings, superior to the pixel-based similarity used in older systems like VIVA.
- Significant performance gains (up to 10.4 times faster than state-of-the-art video analytics systems) while keeping accuracy/diversity.

Weaknesses

- The approach is still weak on relations, counting, and fine-grained attributes. Examples: "dog under table," or "exactly three cars" or "Golden Retriever vs Labrador."
- Synonym and discriminator strategy may struggle on domain-specific vocabularies when label coverage is limited.
- The paper mentions using a fixed 0.80 similarity threshold for diversity; this hard cutoff may not generalize across datasets or queries and could either over-prune or under-prune results.

Discussion

- Zelda uses a fixed 0.80 similarity threshold cutoff. Could thresholds adapt based on query type, dataset, or user signals for better balance between relevance and diversity?