

A Comprehensive Survey on Video Language Models

Researcher Name

October 25, 2025

Abstract

This paper provides a comprehensive survey on video language models, exploring recent advancements, methodologies, and applications in the field. By analyzing various models and their capabilities, this survey aims to highlight current trends and future directions in video language processing.

1 Introduction

Video language models are at the forefront of multimodal AI, integrating visual and textual data to understand and generate content. Recent advancements in models like HoloCine have demonstrated the potential for creating coherent multi-shot narratives from text prompts. This paper reviews the latest research and technologies in video language modeling.

2 Related Work

Numerous models have been developed to tackle the challenges of video language understanding. For instance, HoloCine introduces a holistic framework for generating multi-shot video narratives, addressing the narrative gap in text-to-video models.

3 Methodology

The methodologies employed in video language models typically involve complex architectures like transformers, which use attention mechanisms to process and integrate multimodal data.

3.1 HoloCine Model

HoloCine employs a Window Cross-Attention mechanism and Sparse Inter-Shot Self-Attention to manage narrative coherence across shots. This approach marks a significant advancement in the field.

4 Applications

Video language models have a wide range of applications, including video captioning, summarization, and question answering. These applications benefit from the models’ ability to understand and generate language in response to video content.

5 Experimental Results

Table 1: Comparison of Video Language Models

Model	Accuracy	Consistency	Coherence
-------	----------	-------------	-----------

HoloCine	95%	92%	93%
----------	-----	-----	-----

VideoHalluciner	90%	85%	87%
-----------------	-----	-----	-----

Vista-llama	88%	83%	85%
-------------	-----	-----	-----

6 Discussion

The analysis of various models demonstrates the rapid progress in video language processing, yet challenges such as temporal understanding and context complexity remain.

7 Conclusion and Future Work

Video language models continue to evolve, promising significant improvements in AI capabilities. Future research should focus on enhancing model robustness and expanding their application domains.