WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

VideoSense AI: An Intelligent Multimodal Video Understanding and Analysis System

MS. Keerthana M[1], Harish R[2]

[1] Department of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
[2] Department of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

The rapid growth of digital video has increased the need for systems that can interpret visual content efficiently and present meaningful information to users. Conventional video review often requires manual inspection, which can be time-consuming when recordings are long or contain multiple events. This paper presents VideoSense AI, an intelligent video analysis system designed to support automated video ingestion, frame sampling, object detection, contextual interpretation, natural-language summarization, and notification-based reporting. The proposed workflow accepts common video formats, extracts representative frames at configurable temporal intervals, and applies an open-vocabulary object detection approach using Grounding DINO to identify objects described through text prompts. Detection outputs and frame-level information are organized as structured analysis data and are intended to support higher-level contextual summaries through natural language processing techniques. A notification and reporting interface is included to present analysis outcomes to users. The system is designed as an integrated application rather than an isolated detection model, connecting video processing, model inference, data persistence, and user-facing presentation. The current prototype establishes the processing workflow and application architecture. Quantitative performance claims are not made in this paper because a controlled evaluation dataset and measured benchmark results have not yet been supplied; these are identified as necessary next steps. The work provides a foundation for extensible video understanding, with future improvements focused on detection accuracy, temporal reasoning, model evaluation, and scalable deployment.

Keywords: Video Understanding, Multimodal Analysis, Object Detection, Grounding DINO, Natural Language Processing, Video Summarization

1. Introduction

Video has become a widely used medium for recording activities, documenting events, monitoring environments, and communicating information. The increasing availability of video data has created a practical challenge: extracting relevant information from recordings can require substantial human effort. Reviewing a long video frame by frame is inefficient, and important objects or events may be difficult to locate without automated assistance.

Computer vision techniques can identify visual elements such as people, vehicles, and other objects, while natural language processing techniques can convert structured observations into readable descriptions. However, a useful video analysis application must combine multiple stages, including media ingestion, frame selection, visual inference, result organization, contextual interpretation, and presentation. A detection model alone does not provide a complete video understanding workflow.

VideoSense AI is proposed as an integrated system for processing uploaded videos and presenting analysis information through a unified application. It combines temporal frame sampling with text-guided object detection and a natural-language analysis layer. The system is intended to reduce manual review effort by organizing frame-level detections and producing human-readable summaries and notifications.

The main contributions of the project are: (1) a modular video ingestion and frame extraction workflow for common video formats; (2) integration of Grounding DINO for prompt-based object detection; (3) organization of frame-level analysis outputs for downstream contextual interpretation; (4) a natural-language summarization workflow intended to convert analysis data into concise textual descriptions; and (5) a notification and reporting interface for presenting analysis outcomes.

2. Problem Statement

Manual video inspection is time-consuming and does not scale efficiently when a large number of recordings must be reviewed. Basic automated systems may detect objects in individual frames but may not provide an integrated mechanism for relating detections to the overall video context. In addition, presenting raw bounding boxes and confidence values may not be sufficient for users who need concise explanations of the analyzed content.

The problem addressed in this work is the design and implementation of an integrated application that accepts video files, samples frames, detects user-specified objects, organizes the resulting observations, and presents analysis information in a form that can be understood and reviewed by users. The

system must also preserve the relationship between detections and their corresponding video frames so that results can be traced back to the source recording.

3. Objectives

To develop a video upload and ingestion module that supports commonly used video container formats, including MP4, AVI, MOV, and MKV, subject to the installed decoding libraries.

To extract representative frames using configurable temporal sampling intervals.

To integrate Grounding DINO for text-prompt-based object detection and localization.

To store and organize frame-level detections and associated metadata for subsequent analysis.

To develop an NLP-based workflow that transforms available analysis data into concise, readable video summaries.

To provide a notification and reporting interface that displays analysis outcomes using application data rather than static demonstration content.

To establish an extensible architecture that can support additional models, evaluation procedures, and future multimodal capabilities.

4. Related Work

Object detection has evolved from fixed-category classification systems toward approaches that can identify objects based on language descriptions. Grounding DINO combines a Transformer-based detector with grounded pre-training and supports open-set detection using category names or referring expressions (Liu et al., 2024). This capability is relevant to VideoSense AI because users can specify target objects through text prompts rather than being restricted to a small, predefined list of categories.

The Transformer architecture introduced an attention-based approach to sequence modelling and has influenced a wide range of natural language processing systems (Vaswani et al., 2017). Transformer-based language models provide a foundation for generating and organizing textual descriptions. In the proposed application, natural-language generation is treated as a downstream interpretation stage that uses structured video-analysis information as input.

Research on video summarization has explored methods for selecting representative content and reducing lengthy recordings into shorter descriptions or visual summaries. A survey of multi-view video summarization discusses the growth of video data, existing summarization methods, datasets, and research challenges (Hussain et al., 2021). More recent work on video understanding with large language models examines how language models can act as summarizers and interpreters of video-related information (Tang et al., 2026). These studies motivate the separation of visual analysis from textual interpretation in the VideoSense AI workflow.

The proposed project does not claim to introduce a new object detection architecture or a new language model. Its focus is the application-level integration of video processing, text-guided detection, structured result handling, and natural-language reporting in a single workflow.

5. Proposed System

5.1 System Overview

VideoSense AI follows a modular processing pipeline. A video is submitted through the application interface and passed to the backend for validation and processing. The video is decoded, and frames are extracted at selected time intervals. Each sampled frame is passed to the object detection module together with one or more text prompts. Detection outputs are associated with frame identifiers and timestamps. The resulting records are then made available to the contextual summarization and notification components.

The high-level workflow is: Video Upload → File Validation → Video Decoding → Temporal Frame Sampling → Grounding DINO Inference → Detection Result Structuring → Contextual Text Generation → Notification and Report Presentation.

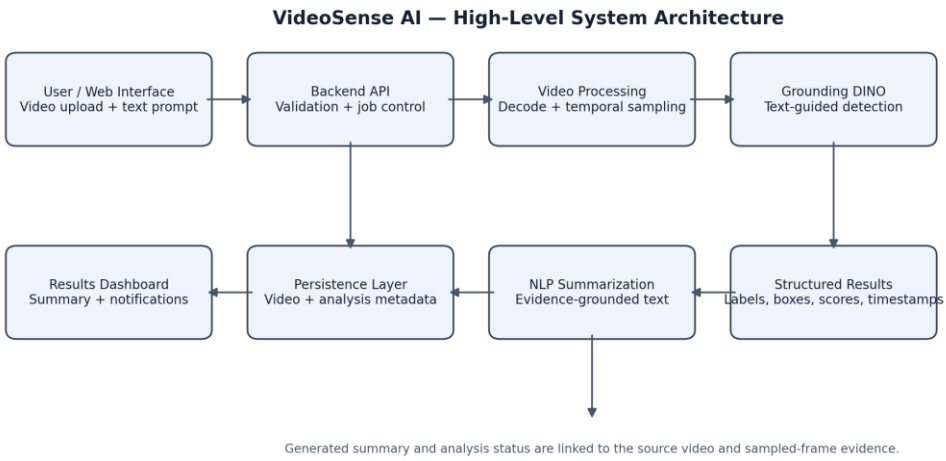


Figure 1. High-level system architecture of VideoSense AI.

5.2 Video Ingestion and Pre-processing

The ingestion module receives a video file and validates its availability, file extension, and processability before analysis begins. The system is designed to support commonly used formats such as MP4, AVI, MOV, and MKV. Actual format support depends on the video-decoding libraries and codecs available in the deployment environment.

After validation, the video is decoded and representative frames are extracted. Temporal sampling is used to avoid processing every frame in a high-frame-rate recording. For a sampling interval of Δt seconds, frames are selected at regular time points across the video duration. This reduces the number of images passed to the detection model and provides a configurable trade-off between processing cost and temporal coverage.

The sampling interval must be selected carefully. A large interval can reduce computation but may miss short-lived events, while a small interval provides denser temporal coverage at a higher computational cost. The selected interval and the timestamp of each extracted frame should be retained as analysis metadata.

5.3 Object Detection Using Grounding DINO

Grounding DINO is used as the visual detection component. It supports open-set object detection by associating image features with text descriptions, enabling detection queries to be expressed in natural language (Liu et al., 2024). For each sampled frame, the system supplies the image and the configured text prompt to the model. The model returns candidate object regions and associated confidence information.

Each detection should be represented with at least the frame identifier, timestamp, queried label, predicted label or phrase, bounding-box coordinates, and confidence score, where those fields are returned by the implementation. A configurable confidence threshold can be applied to reduce low-confidence detections. Threshold selection affects the balance between missed detections and false positives and should be validated against a labelled dataset.

Grounding DINO is an open-vocabulary detection model, but its outputs are not guaranteed to be correct for every scene, object scale, lighting condition, camera angle, or occlusion level. The project therefore treats detections as model predictions rather than verified facts. User review and quantitative evaluation remain important for reliability.

5.4 Contextual Analysis and NLP-Based Summarization

Frame-level detections provide localized information but do not automatically describe the complete video. VideoSense AI therefore organizes detections and timestamps into structured analysis records before passing relevant information to a natural-language summarization stage. The summarization component is intended to convert the available records into a concise description of detected objects and their observed distribution across sampled frames.

The summary-generation input should be restricted to information actually produced by the analysis pipeline. For example, the presence of a detected object in a particular frame can be described when supported by a detection record. Claims about intent, identity, causality, or events that are not represented by the available visual evidence should not be generated as confirmed statements. This evidence-grounded approach helps reduce unsupported descriptions.

Where multiple detections refer to the same object category across nearby timestamps, the system may aggregate them into a temporal overview. More advanced tracking and event reasoning can be incorporated in future versions, but should only be reported as implemented after those capabilities have been integrated and tested.

5.5 Notification and Reporting Module

The notification and reporting module presents the results of completed video analyses. It is intended to retrieve real analysis records, including processing status, video reference, timestamps, detection summaries, and generated textual descriptions. Notifications should be created from backend events or persisted analysis states rather than hard-coded sample messages.

The interface can provide a summary of the analyzed video, selected detected objects, processing status, and links or references to relevant frames. Error states, incomplete processing, and failed inference should be represented separately from successful analysis results so that users are not shown misleading completion messages.

5.6 Data Management and Application Architecture

The application is organized into user-facing and backend responsibilities. The frontend provides video submission and result visualization. The backend manages file validation, frame extraction, model inference, result processing, and communication with the persistence layer. Analysis records should maintain a stable association with the uploaded video and the sampled frame from which each detection originated.

A modular architecture allows the detection model, summarization component, and notification mechanism to be updated independently. It also supports future integration of additional modalities, such as audio transcription, provided that audio processing is implemented and evaluated separately. The term multimodal in this paper refers to the intended integration of visual and language information; audio analysis should not be assumed unless it is part of the deployed implementation.

6. Implementation Methodology

6.1 Processing Workflow

Video submission: The user selects and uploads a supported video file through the application.

Validation: The backend checks the file and confirms that it can be processed.

Frame extraction: The video is decoded and frames are sampled at the configured temporal interval.

Prompt-based inference: Each sampled frame is submitted to Grounding DINO with the selected object prompt.

Result normalization: Model outputs are converted into structured records with frame references and timestamps.

Summary generation: The NLP component uses the structured analysis records to produce a concise textual overview.

Notification and display: The application presents processing status, detections, and the generated summary.

6.2 Technology Components

The available project documentation confirms a Python-based backend and integration work involving Grounding DINO. The development environment also uses PyTorch, as indicated by the reported PyTorch runtime warning during model setup. However, the supplied manuscript does not contain the dependency lockfile, package manifest, frontend package configuration, database configuration, or NLP provider/model identifier. Consequently, exact version numbers and several implementation choices cannot be verified from the materials reviewed. Table 1 separates confirmed components from details that must be read from the deployed source code before submission.

Layer / Component	Status supported by available evidence	Version / submission action
Backend language	Python-based backend is confirmed.	Exact Python version: extract from `python --version` in the project virtual environment.
Deep-learning runtime	PyTorch is confirmed by the reported runtime warning.	Exact version: extract using `python -c "import torch; print(torch.__version__)"`.
Object detection	Grounding DINO integration is identified.	Record repository commit/checkpoint variant and configuration; do not treat paper publication year as software version.

Layer / Component	Status supported by available evidence	Version / submission action
Video decoding / frame extraction	A media-processing dependency is used, but its identity is not stated in the manuscript.	Confirm package name and version from imports and dependency files (for example, do not assume OpenCV without checking).
Frontend framework	A web application interface is confirmed; framework is not identified.	Enter the framework and version from `package.json` or the actual frontend configuration.
Database / persistence	A persistence layer is described conceptually; database product is not identified.	Enter database engine and version from application configuration.
NLP summarization	An NLP summarization stage is described; model/provider is not identified.	Enter exact model name, provider, API/model version, and prompting method.
Operating system / hardware	Not recorded in the supplied manuscript.	Record OS, CPU, GPU, VRAM, RAM, and CUDA version where applicable.

Note. The version commands above are verification instructions, not reported experimental values. Replace each action with the exact environment output before journal submission.

6.3 Data Flow and Traceability

For reproducibility, every analysis should retain a unique analysis identifier and a reference to the original video. Extracted frames should be associated with timestamps, and detection records should refer to the corresponding frame. Generated summaries should be linked to the analysis record from which they were produced. This data flow makes it possible to inspect the evidence supporting a summary and to diagnose incorrect detections.

7. Evaluation Methodology

A research evaluation is required to determine whether VideoSense AI provides reliable object detection and useful summaries. The evaluation should use a representative collection of videos with manually annotated object labels and, where relevant, timestamps and bounding boxes. The dataset should include different object categories, camera viewpoints, lighting conditions, object sizes, occlusion levels, and scene types.

7.1 Object Detection Metrics

Object detection performance should be measured using precision, recall, F1-score, and mean Average Precision (mAP), where suitable ground-truth bounding-box annotations are available. Intersection over Union (IoU) should be used to determine whether a predicted bounding box sufficiently overlaps with a ground-truth box. The confidence threshold and IoU threshold must be stated clearly in the final experimental report.

7.2 Processing Performance

Processing performance should be measured using total analysis time, frame extraction time, average inference time per sampled frame, number of sampled frames, and resource consumption where available. Tests should record the hardware configuration, model variant, input resolution, sampling interval, and software environment so that the results can be interpreted and reproduced.

7.3 Summary Quality

Generated summaries should be assessed for factual consistency with the underlying detection records, coverage of relevant observations, readability, and conciseness. Human evaluation by multiple reviewers is recommended. If reference summaries are prepared, automatic text-similarity metrics may be reported as supplementary measures, but they should not replace evidence-based human assessment.

8. Discussion

The principal value of VideoSense AI is its integrated workflow. By connecting video ingestion, temporal sampling, text-guided object detection, structured data handling, natural-language summarization, and notification presentation, the system provides a foundation for reducing the effort required to review video content. The use of text prompts also provides flexibility when users need to search for different object categories.

The quality of the final analysis depends on each stage of the pipeline. Inaccurate frame sampling can omit important visual information. Detection errors can propagate into the structured records, and unsupported language generation can make an otherwise correct detection pipeline appear unreliable. Consequently, the summarization component should remain grounded in model outputs, and users should be able to inspect the frame-level evidence.

A further limitation is that object detection is not equivalent to full event understanding. Identifying an object in multiple frames does not, by itself, establish its actions, relationships, or intent. Temporal tracking, action recognition, audio analysis, and event reasoning would require additional models and evaluation. These capabilities are considered extensions rather than confirmed features unless they are present in the final implementation.

9. Limitations and Future Enhancements

Dataset-specific fine-tuning and validation may be required to improve detection performance for objects and environments that are not adequately represented in the base model's training data.

Object tracking can be added to associate detections across consecutive frames and support more reliable temporal descriptions.

Action recognition and event detection can be integrated to move beyond frame-level object presence toward activity-level interpretation.

Audio transcription and audio-visual alignment can be explored to support a broader multimodal analysis pipeline.

Summary generation can be improved through evidence-grounded prompting, structured output validation, and human review.

Scalability can be improved through asynchronous processing, job queues, optimized inference, and resource-aware frame sampling.

Security and privacy controls, including upload validation, access management, retention policies, and secure storage, should be assessed before production deployment.

10. Conclusion

This paper presented VideoSense AI, an integrated application concept for intelligent video analysis and reporting. The system combines video ingestion, temporal frame extraction, Grounding DINO-based text-guided object detection, structured analysis records, natural-language summarization, and notification-oriented presentation. The design emphasizes traceability between the original video, sampled frames, detection outputs, and generated summaries. The current project establishes a foundation for an extensible video understanding workflow; however, reliable claims about accuracy, speed, and summary quality require controlled experiments using annotated data. Future work will focus on model validation and fine-tuning, temporal reasoning, improved evidence-grounded summaries, and scalable deployment.

REFERENCES

Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S. W., & de Albuquerque, V. H. C. (2021). A comprehensive survey of multi-view video summarization. *Pattern Recognition, 109*, Article 107567. https://doi.org/10.1016/j.patcog.2020.107567

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2024). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024* (pp. 38–55). Springer. https://doi.org/10.1007/978-3-031-72970-6_3

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Gaudiot, J.-L., Luo, P., Luo, J., & Xu, C. (2026). Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology, 36*(2), 1355–1376. https://doi.org/10.1109/TCSVT.2025.3566695

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html