Proceedings of Machine Learning Research 288:1–17, 2025 2nd International Conference on Neuro-symbolic Systems (NeuS)

A Challenge to Build Neuro-Symbolic Video Agents

Sahil Shah SS96869@UTEXAS.EDU
Harsh Goel HARSHG99@UTEXAS.EDU
Sai Shankar Narasimhan NSAISHANKAR@UTEXAS.EDU
Minkyu Choi MINKYU.CHOI@UTEXAS.EDU
S P Sharan SPSHARAN@UTEXAS.EDU
Oguzhan Akcin OGUZHNAKCIN@UTEXAS.EDU
Sandeep Chinchali SANDEEPC@UTEXAS.EDU
The University of Texas at Austin

Editors: G. Pappas, P. Ravikumar, S. A. Seshia

Abstract

Modern video understanding systems excel at tasks such as scene classification, object detection, and short video retrieval. However, as video analysis becomes increasingly central to real-world applications, there is a growing need for *proactive* video agents—systems that not only interpret video streams but also reason about events and take informed actions. A key obstacle in this direction is *temporal reasoning*: while deep learning models have made remarkable progress in recognizing patterns within individual frames or short clips, they struggle to understand the sequencing and dependencies of events over time, which is critical for action-driven decision-making. Addressing this limitation demands moving beyond conventional deep learning approaches. We posit that tackling this challenge requires a *neuro-symbolic* perspective, where video queries are decomposed into atomic events, structured into coherent sequences, and validated against temporal constraints. Such an approach can enhance interpretability, enable structured reasoning, and provide stronger guarantees on system behavior, all key properties for advancing trustworthy video agents. To this end, we present a grand challenge to the research community: developing the next generation of intelligent video agents that integrate three core capabilities—(1) autonomous video search and analysis, (2) seamless real-world interaction, and (3) advanced content generation. By addressing these pillars, we can transition from *passive* perception to intelligent video agents that reason, predict, and act, pushing the boundaries of video understanding.

Keywords: Video Understanding · Neuro-Symbolic AI · Video Agent · Video Generation

1. Introduction

Consider an agentic system that utilizes video feed from the cameras of a home security system (Figure 1) to perform the following complex user request:

> *Send me a notification when a person is walking up to my driveway with a package, then automatically open the garage for three minutes to safely store the package and lock everything back up when the delivery is complete.*

Home-security owners increasingly expect video camera systems to handle tasks of this complexity (Comeau, 2024). Beyond delivery automation, such a system should also be able to notify law enforcement authorities if it recognizes an attempted break-in, send the exact video footage of the incident, and autonomously trigger an alarm to lock all entry points to secure the premises.

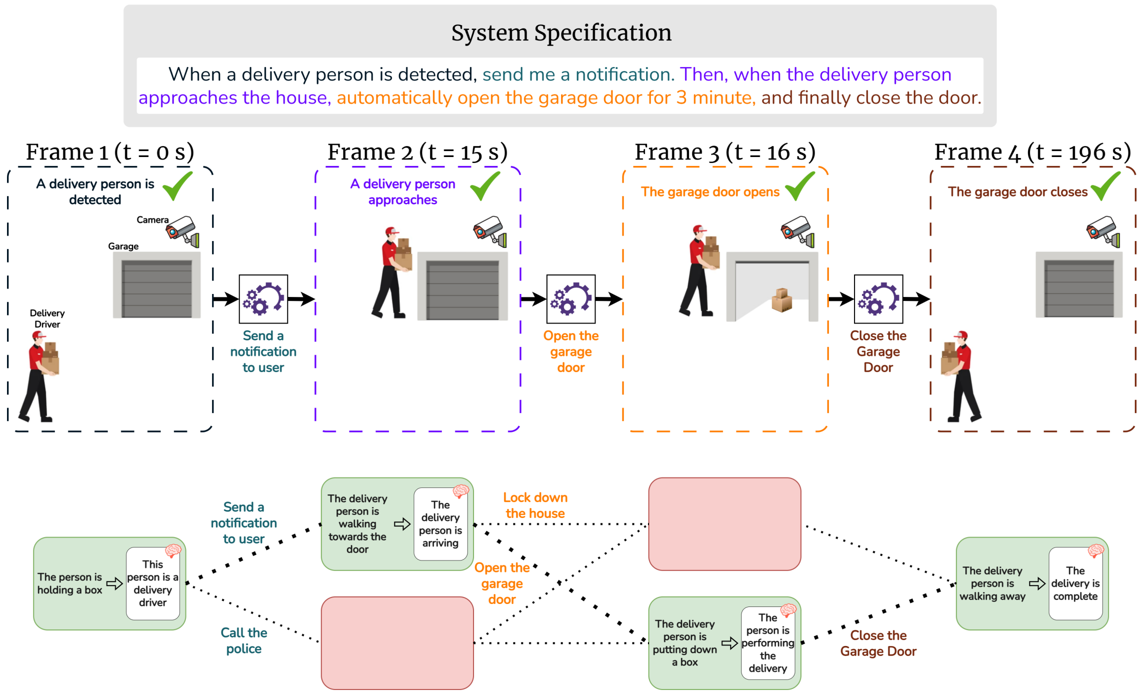


Figure 1: **An Efficient Neuro-Symbolic Approach to Video Agents.** We argue for a neuro-symbolic approach to develop video agents that combines the per-frame or short-horizon reasoning capabilities of neural perception models with the long-term reasoning abilities of symbolic frameworks such as temporal logic tools. Here, we show one such example from a home security system, where the agent is required to identify the presence of a delivery person and send the required notifications.

Requests of this nature are not limited to home security; similar demands arise in defense (Shultz and Clarke, 2020), autonomous driving (National Highway Traffic Safety Administration, 2024), and social media analytics (Metricom, 2025). Endowing current video systems to process user queries as shown above requires moving beyond *passive* analysis to *reason* about unfolding events in real-time, and precisely *interact* with the real world with guarantees.

While deep learning excels at understanding short-term activities (*e.g.*, detecting a person, object, or short event), it struggles with temporal reasoning and long-term memory (see Section 2), posing a significant challenge for video agents tasked with understanding complex user queries (Choi et al., 2024). Furthermore, deep learning methods in isolation do not provide the necessary interpretability or guarantees for either perceiving or acting.

In contrast, structured logic representations, such as temporal logic (TL) (Baier and Katoen, 2008), effectively address these challenges by explicitly encoding time-dependent sequences of events such as "close the garage door within 20 minutes of opening it." Therefore, we argue that the next generation of video understanding systems will adopt a hybrid of deep learning (*neuro*) and formal representations such as TL (*symbolic*). A **neuro-symbolic** approach can capture complex temporal constraints while providing interpretability and formal guarantees with respect to user specifications. Therefore, we posit that future video systems—which we term **video agents**—will be built on three pillars:

1. **Video Search and Understanding**: A video agent must identify and extract the relevant frames that correspond to a complex temporally extended query. For instance, in the running example, tagging a person with a package walking towards the front door as an attempted delivery would require the system to process the progression of events over time.

2. **Integrate Understanding with Real-World Action:** Video agents must act in the real world according to user specifications by integrating perception with downstream actions such as calling relevant Application Programming Interfaces (APIs) (Chase, 2022). For example, actions such as "Opening the garage for three minutes after detecting a delivery" or "Locking down entry points during an attempted break-in" are enacted only after specific sequences of events are identified.

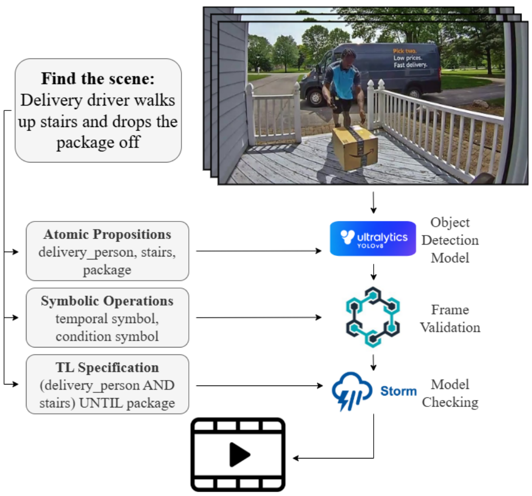
3. **Video Generation:** Video agents must be rigorously tested through synthetic simulations for edge cases, for instance, distinguishing between a delivery driver and an intruder or detecting suspicious packages. Additionally, video generation serves as a tool for these agents to create video clips that edit out sensitive user content for post-hoc analysis, for instance, in our example, generating or editing video clips for law enforcement or delivery companies.

Hence, we propose a grand vision and challenge for the community to develop the next generation of neuro-symbolic video agents capable of analyzing videos (both offline and in real time) and executing user-specified actions. We outline objectives for each desired capability—video search, action execution, and video generation—and provide a starting point with relevant datasets (TLV (Choi et al., 2024) for video search) and metrics (NeuS-V (Sharan et al., 2024) for video generation) to motivate further research and development in this direction. Finally, we urge the research community to establish rigorously standardized benchmarks and systematically validate video agents as these agents transition into safety-critical domains such as autonomous driving and home security.

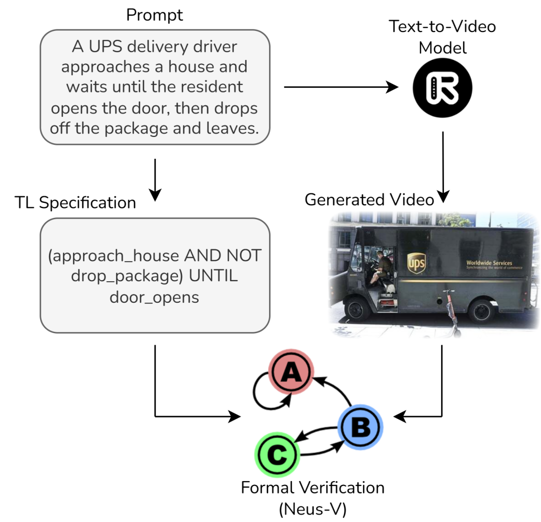
2. Can Deep Learning Alone Solve Video Understanding?

In this section, we revisit a fundamental question: Can deep learning alone achieve comprehensive video understanding, the key building block for any video agent? While state-of-the-art multimodal foundation models such as GPT4 (Achiam et al., 2023), LLaMA (Touvron et al., 2023), and Gemini (Team et al., 2023) have demonstrated impressive capabilities for language and image tasks, these models lack the ability to interpret extended temporal dependencies between events. In the following, we highlight the key drawbacks of these foundation models in the context of video search and generation, both critical for the agent's operation.

Traditional deep learning models struggle to capture complex and long-term temporal dependencies in videos. We attribute this to the temporal aggregation of semantic and activity-related deep learning, which couples spatial and temporal feature processing. This limitation is demonstrated with Neuro-Symbolic Video Search with Temporal Logic (NSVS-TL) (Choi et al., 2024). NSVS-TL, as illustrated by its pipeline in Figure 2(*a*), maps videos into a probabilistic automaton by converting frames into states using an off-the-shelf neural perception model. Complex user queries are then converted into TL specifications. Consequently, NSVS-TL converts the video search problem into a verification problem and extracts the relevant clips that satisfy the TL specifications corresponding to the user query. By decoupling spatial and temporal understanding, this approach



(*a*) NSVS-TL Architecture.



(*b*) NeuS-V Architecture.

Figure 2: **NSVS-TL and NeuS-V System Diagrams.** In (*a*), when given video feed from a security system, NSVS-TL demonstrates its capability in identifying exactly when a delivery driver walks up the stairs and drops a package off. Similarly, in (*b*), a video generated by a foundation model describing a delivery scenario is evaluated for temporal fidelity through NeuS-V.

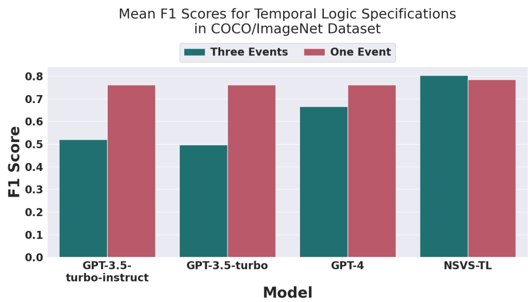
significantly outperforms multimodal foundation models when faced with elaborate user queries, as shown in Figure 3(*a*).

Foundation models for video generation deteriorate in performance with increasing complexity of the text prompts. Dozens of text-to-video generation models exist (see Section 3 for a summary), however, evaluations on benchmarks such as NeuS-V (Sharan et al., 2024), with its pipeline depicted in Figure 2(*b*), demonstrate a deficiency in temporal fidelity, where the semantic alignment between video and prompt deteriorates with increasing user query complexity. Inspired by NSVS-TL, NeuS-V first translates the text prompt into atomic propositions and a TL specification using an optimized large language model (LLM). It then constructs a video automaton representation by assigning semantic confidence scores to these propositions using a vision-language model (VLM). Finally, NeuS-V computes the satisfaction probability by formally verifying the video automaton against the TL specification to produce the final score. In Figure 3(*b*), we show that text-to-video generation degrades when the number of events in the query increases. To mitigate this issue, we posit that generative models need to be co-designed with temporal understanding frameworks.

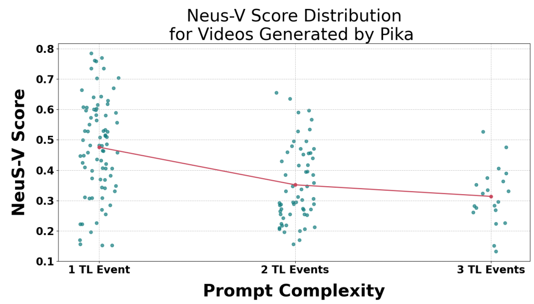
Overall, deep learning models still struggle with both long-form video understanding and video generation. Hence, video agents would need to be built via neuro-symbolic methods that blend deep learning methods with formal logic representations and state machines.

3. Related Work

Video Search Existing research in video event detection predominantly focuses on tracking spatio-temporal object information using deep neural networks to learn latent representations, such as mo-



(*a*) NSVS-TL excels at complex video search.



(*b*) Video generation fails with complexity.

Figure 3: **Foundation models struggle to perform video search and generation with increasing complexity of user queries.** However, neuro-symbolic approaches (NSVS-TL) effectively decouple spatial and temporal reasoning using perception modules for spatial reasoning and temporal logic (TL) to model long-term temporal dependencies. As a result, NSVS-TL outperforms foundation models in complex video search tasks (*a*). Similarly, text-to-video models, like Pika, fail to maintain temporal consistency as scenario complexity increases (*b*).

tion and position, for event detection and classification (Jiang et al., 2011; Medioni et al., 2001; Li et al., 2022; Xu et al., 2015; Zheng et al., 2022; Doshi and Yilmaz, 2023; Feichtenhofer et al., 2016, 2017). These methods, while effective, demand substantial computational resources for training and inference (Li et al., 2022; Feichtenhofer et al., 2016, 2017). Some works also extend this to natural language-based event detection, employing VLMs like Video-LLaMA (Zhang et al., 2023a) and Video-ChatGPT (Maaz et al., 2023), which integrate language foundation models such as GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023), for zero-shot event recognition and visual question answering (VQA). However, their reliance on temporal aggregation limits precise frame identification in long videos. Neuro-symbolic methods have been proposed to address this issue by enabling structured reasoning over longer videos when searching for video clips (Cheng et al., 2014; Yang et al., 2023; Choi et al., 2024, 2025b).

Video Agents Recent advancements in video agents have leveraged LLMs and VLMs to enable decision-making, such as ReACT (Yao et al., 2023) powered through Langchain (Chase, 2022) and ToolLLM (Qin et al., 2023). These have been applied in the video domain where tools have been integrated with VLMs for tasks such as long-from video understanding (Jeoung et al., 2024; Wang et al., 2024b), video generation for robotics (Soni et al., 2024), and video editing (Wang et al., 2024a). While these agents demonstrate progress, they cannot be applied in real-world applications, such as home security, which require the seamless integration of temporal reasoning, context-aware decision-making, and robust interaction with external systems.

Video Generation Following the recent successes in text-to-image generation, text-to-video models such as SORA from OpenAI (OpenAI, 2024), GEN-3 Alpha from Runway (Research, 2024), and PIKA (Labs, 2024) have seen increased proliferation. Although the exact architecture for these models is unknown, text-to-video generation employs diffusion models (Ho et al., 2020), such as in Phenaki (Villegas et al., 2022) and I2VGen-XL (Zhang et al., 2023b; Blattmann et al., 2023; Esser et al., 2023), or autoregressive models, such as the CogVideo series (Hong et al., 2022; Yang et al., 2024). For a detailed survey on these methods, we refer you to Cho et al. (2024). These models are inadequate for generating temporally correlated events (Sharan et al., 2024; Choi et al., 2025a), and to our knowledge, neuro-symbolic methods for solving this problem have not been explored.

Neuro-symbolic Methods Many works explore approaches to building symbolic representations for video classification (Feichtenhofer et al., 2019; Tran et al., 2019), event detection (Medioni et al., 2001; Xu et al., 2015; Li et al., 2022), video question-answering (Yi et al., 2018; Chen et al., 2022), robotics (Shoukry et al., 2017; has, 2019; Kress-Gazit et al., 2009), and autonomous driving (Jha et al., 2018; Mehdipour et al., 2023). These methods either construct graph structures (Yu et al., 2022; Mavroudi et al., 2020; Xiong et al., 2019), use latent-space representations as symbolic representations (Sarkar et al., 2015; Bertasius et al., 2021; Kroshchanka et al., 2021), or leverage formal language methods (Baier and Katoen, 2008) to design specifications.

4. The Formal Challenge

Our goal is to encourage the research community to develop the analogous version of the LLM-based agent datasets, tasks, and evaluation frameworks (*e.g.*, Tool Bench (Qin et al., 2023), Stable Tool Bench (Guo et al., 2024), and Gorilla (Patil et al., 2023)) for the video domain to create and evaluate video agents. To this end, we formalize the design goals of a video agent that leverages deep-learning and neuro-symbolic methods to process videos and complex natural language queries. The inputs and outputs of such a system are:

- **Inputs**
  - **Dataset:** Videos annotated with natural language queries for temporally complex events and their corresponding ground truth spans or actions. We provide a preliminary dataset used for video search, the TLV Dataset in Section 5, for the community to build upon and adapt for video agents.
  - **Tools**: A set of apps or programs to be executed, such as calling Python code, executing a state machine, or calling an API from an external library like Twilio (Twilio Inc., 2025) and RapidAPI (RapidAPI). We encourage the community to develop open-sourced API's specific to video agent use-cases.
  - **Models:** Deep learning models such as LLMs, VLMs, and Generative Models.
- **Outputs**
  - **Event Clips:** A sequence of specific frames that correspond directly to a user query.
  - **Actions:** A tool invoked with its inputs timed according to the user query.
  - **Synthetic Videos:** Synthesized videos based on user queries and specifications.
- **Metrics**
  - **Accuracy of Events:** F1-Score between the predicted spans and ground-truth spans.
  - **Tool Calling:** Accuracy of the selected tool and its desired input (see Section 6).
  - **Synthetic Videos:** VBench (Huang et al., 2024) for visual quality and NeuS-V (Sharan et al., 2024) for temporal fidelity.

The primary objective of the video agent is to merge agentic workflows capable of tool invocation (Chase, 2022) based on input videos and queries. Therefore, given a dataset annotated with events and corresponding tool actions, the agent would predict which tool should be invoked for each query at specific instances in the video. The core technical requirements for each of the capabilities for the video agent are:

1. **Video Search:** The video search task aims to predict the temporal span of a video clip that corresponds to a given natural language query. This requires:
   (a) **Parsing Queries:** Decompose the natural language query into semantic components that include objects and short-term events or activities followed by a formal language description of their temporal order.
   (b) **Perception:** Orchestrate neural models to detect relevant objects and activities.
   (c) **Prediction:** Predict temporal spans in the query-aligned video with high probability.
2. **Tool Calling:** The key capability of the video agent is to be able to call the right tools and specify their input based on the user's specifications and the context developed from the video. The implementation of this task requires:
   (a) **Tool Selection**: Run deep learning models to identify the right tool or API to call based on the identified temporal spans from the video search task.
   (b) **Tool Invocation:** Specify the inputs to the tool to be executed based on the identified video clip—for instance, the message to be sent for a notification API.
3. **Video Generation:** The video generation task aims to synthesize videos that align with a temporally extended natural language query. To accomplish this, the following are necessary:
   (a) **Synthesis:** Synthesize novel scenarios based on the user's custom queries.
   (b) **Evaluation:** Ensure high visual quality while maintaining semantic coherence and accurate temporal alignment with the query.
   (c) **Improvement and Editing:** Iteratively improve the videos through reprompting or editing from neuro-symbolic feedback to meet the user's specifications.

We provide initial data along with benchmarks and example agents for the community to build upon on GitHub.

5. Datasets

A plethora of video datasets, such as Ego4D (Grauman et al., 2022), MSR-VTT (Xu et al., 2016), and others (Nagrani et al., 2022; Chandrasegaran et al., 2025; Wang et al., 2023), annotate short video clips with activities in natural language. However, **they are not suitable for our purpose** because they lack annotations for temporally structured activities such as "The garage door opens after the person is identified as a delivery driver."

To develop video agent capabilities such as search, generation, and tool invocation, datasets must include three key annotations: (1) frame-level temporal annotations for events (*e.g.*, detecting deliveries or break-ins), (2) specifications of tools and their inputs (*e.g.*, notification type and recipient), and (3) the temporal order of tool invocations (*e.g.*, "alert homeowner before locking doors").

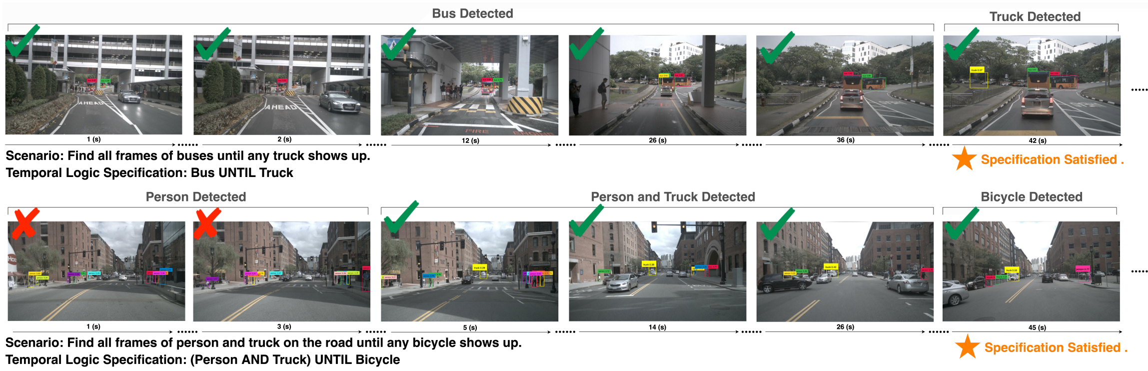


Figure 4: **TLV Dataset Specification-Video Pairing.** An excerpt of the TLV dataset is shown here, demonstrating the efficacy of the TLV dataset in showing the relationships between videos and TL sequences. This figure was taken with permission from Choi et al. (2024).

The TLV dataset is a first step towards addressing these limitations. The TLV dataset (Choi et al., 2024) recognizes the need to explicitly annotate *when* an event occurs and how they are *temporally related*. At a high level, the TLV dataset is designed with frame-level temporal annotations for temporally dependent events. These annotations are compiled from a combination of static images from leading image datasets, including Waymo (Sun et al., 2020) and NuScenes (Guo et al., 2020), as illustrated in Figure 4, and the dataset is publicly accessible through Hugging Face. While TLV facilitates evaluations on video content for video search with temporally correlated event queries, it lacks the annotations pertaining to the tool invocation. This poses an interesting challenge, and we empower the community to curate temporally meaningful datasets with timestamped annotations of the tools used on the more modern video-activity datasets listed above.

6. Metrics

Consider our running example described in Section 1.

> *How can we evaluate the success of a video agent that is capable of video search and toolchain execution in both real-world and good quality synthetically generated scenarios?*

A comprehensive evaluation metric for agents would consider computational efficiency, video processing latency, state machine generation latency, and resource consumption, such as energy and memory. However, for this challenge, we will focus specifically on evaluating the accuracy of the video agent and video generation.

Event Search Accuracy for Video Search In the video search domain, we propose utilizing classification metrics, such as the F1 score. This metric, the aggregation of precision and recall, is calculated based on the accurate identification of frames relevant to a given search query compared against ground truth labels provided by datasets like the TLV dataset.

Evaluating Tool Calling A comprehensive metric to evaluate video agents would ensure that agents interpret events accurately, maintain temporal coherence with user specifications, and execute actions with appropriate inputs. We list the components as follows:

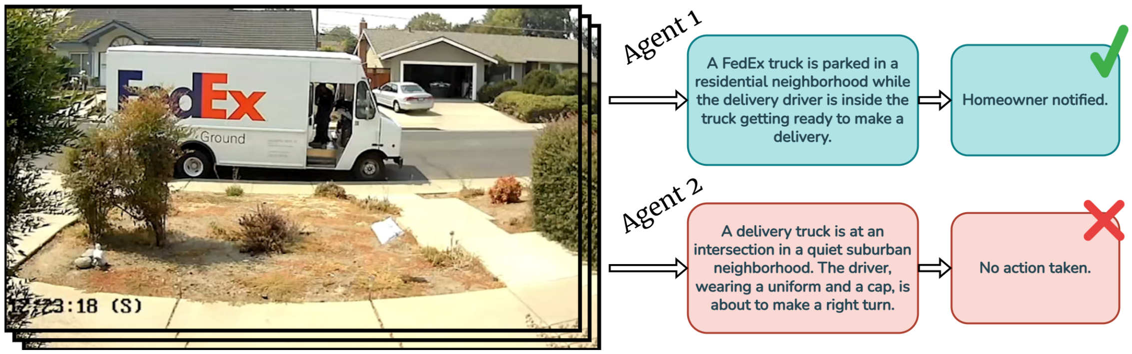


Figure 5: **What is the Correct Action for the Agent?** At a first glance, both agents summarize the video nearly identically. However, upon closer inspection, Agent 1, although more vague, correctly identifies the parked delivery truck and notifies the homeowner. In contrast, Agent 2, while more detailed, hallucinates a right turn on an intersection, leading to no action.

- **Event-Specific Tool Invocation:** The metric must assess the accuracy of the tools executed in response to specific events. For instance, in Figure 5, the agent must correctly notify the homeowner when a delivery person is detected approaching.
- **Temporal Alignment with the Prompt:** The evaluation must ensure that the temporal order of tool invocation adheres to the specified sequence. For example, the garage door should be closed only after the delivery person leaves the driveway but before they leave the package. Similarly, if a break-in is detected, the system must simultaneously trigger an alarm, lock all entry points, and notify law enforcement authorities.
- **Correct Inputs to Tools:** The tools invoked by the agent must receive accurate and contextually appropriate inputs. For instance, notifications should be sent to the correct recipient (*e.g.*, notifying law enforcement in case of a break-in).

Video Generation Evaluation Video generation necessitates an assessment of both visual quality and temporal fidelity. Existing evaluation metrics, such as those based solely on VQA (Wu et al., 2023a,b), emphasize visual quality, neglecting temporal coherence in the process. Consequently, these metrics may yield high scores, even for videos that fail to represent the prompt's intended sequence of events. To address this deficiency, we propose a two-pronged evaluation approach: one benchmark to assess visual quality and another to evaluate temporal adherence. Specifically, we suggest employing VBench (Huang et al., 2024) as the standard for visual quality assessment, complemented by NeuS-V (Sharan et al., 2024), a metric we have discussed earlier that conducts temporal coherence evaluations on the generated video by converting its natural language prompt into atomic events and semantics with TL specifications.

7. Architecture Discussion and Open Questions

In this section, we pose several questions and ideas to the research community to expand the scope for neuro-symbolic video agents to multimodal inputs, multiple formal language representations, and multiagent setups.

Multimodal: Audio, Video, and Beyond. At its core, our description of a video agent decomposes user queries into fundamental atomic events that can be detected using computer vision. However, consider the case where a video agent must respond to multimodal triggers, such as a home security system awaiting the audio of a door knock. How can I modify my agent framework to allow for complex, multimodal triggers? For this system, the role of the neural model extends beyond pure video analysis, encompassing the complex orchestration of multimodal information to facilitate robust and adaptable symbolic reasoning.

How is Formal Language Specified? The temporal structure of events through formal language presents a unique challenge. From Section 1, we described formal language through TL, allowing the video agent to capture temporal information in user queries. However, video agents are not married to a single type of formal language framework; if precise timings were vital, video agents can leverage approaches such as Linear Temporal Logic (LTL) or Signal Temporal Logic (STL). Now, consider the need to track spatial relationships or distances between objects. This idea warrants the exploration of more flexible formal language, such as state machines or pipelines akin to LangChain (Chase, 2022). Ultimately, any approach must be both expressive and scalable, capable of capturing complex temporal dependencies while remaining computationally feasible for verification. Further exploration is encouraged to determine which approach, or combination of approaches, yields the most robust representation framework.

Multi-Agent Equals Multi-Camera? Consider an agentic security system developed for a large apartment complex with dozens of security cameras instead of one. How can these cameras interface with each other to create one cohesive system? This now becomes a multi-agent problem where atomic propositions can be derived from different security cameras across the complex. This introduces several key challenges: What mechanisms facilitate the fusion of disparate spatial and temporal information into a unified representation? Furthermore, how can the system reason about events occurring across multiple viewpoints, effectively resolving potential ambiguities and inconsistencies? The goal for a multi-agent system is to move beyond the simple aggregation of camera data and towards a truly robust system that can reason about complex, distributed events with enhanced situational awareness.

8. Conclusion

The development of neuro-symbolic video agents marks the next frontier in video understanding, blending the pattern recognition capabilities of deep learning with the interpretability and temporal reasoning strengths of symbolic methods. These hybrid systems address critical limitations of current models, such as poor long-term memory and a lack of guarantees for perception and action. By bridging low-level recognition with high-level logic, neuro-symbolic approaches enable robust performance in complex settings like home security, autonomous driving, and beyond. They empower systems to reason over temporal sequences, generate synthetic content for edge cases, and translate understanding into action seamlessly. We encourage the research community to expand this paradigm by exploring architectures and methods that tightly integrate learning with formal reasoning, paving the way towards intelligent video agents.

Acknowledgments

This material is based upon work supported in part by the Office of Naval Research (ONR) under Grant No. N00014-22-1-2254. Additionally, this work was supported by the Defense Advanced Research Projects Agency (DARPA) contract DARPA ANSR: RTX CW2231110. Approved for Public Release, Distribution Unlimited.

References

Reinforcement learning for temporal logic control synthesis with probabilistic satisfaction guarantees. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 5338–5343. IEEE, 2019.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. The MIT Press, 2008.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2025.

Harrison Chase. Langchain, 2022. URL https://github.com/langchain-ai/langchain.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*, 2022.

Yu Cheng, Quanfu Fan, Sharath Pankanti, and Alok N. Choudhary. Temporal sequence modeling for video event detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2242, Columbus, OH, USA, 2014. IEEE Computer Society.

Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024.

Minkyu Choi, Harsh Goel, Mohammad Omama, Yunhao Yang, Sahil Shah, and Sandeep Chinchali. Towards neuro-symbolic video understanding. In *European Conference on Computer Vision*, pages 220–236. Springer, 2024.

Minkyu Choi, SP Sharan, Harsh Goel, Sahil Shah, and Sandeep Chinchali. We'll fix it in post: Improving text-to-video generation with neuro-symbolic feedback. *arXiv preprint arXiv:2504.17180*, 2025a.

Minkyu Choi, Yunhao Yang, Neel P Bhatt, Kushagra Gupta, Sahil Shah, Aditya Rai, David Fridovich-Keil, Ufuk Topcu, and Sandeep P Chinchali. Real-time privacy preservation for robot visual perception. *arXiv preprint arXiv:2505.05519*, 2025b.

Zachary Comeau. Parks research shows consumers are after integrated smart locks and security cameras. *CEPro*, September 2024. URL https://www.cepro.com/news/consumers-want-integrated-smart-locks-and-security-cameras/142262/.

Keval Doshi and Yasin Yilmaz. Towards interpretable video anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2654–2663, Waikoloa, HI, USA, 2023. IEEE.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, Las Vegas, NV, USA, 2016. IEEE Computer Society.

Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7445–7454, Honolulu, HI, USA, 2017. IEEE Computer Society.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 6201–6210. IEEE, 2019.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

Yiluan Guo, Holger Caesar, Beijbom Oscar, Jonah Philion, and Sanja Fidler. The efficacy of neural planning metrics: A meta-analysis of pkl on nuscenes. In *IROS 2020 Workshop on Benchmarking Progress in Autonomous Driving*, 2020.

Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive video understanding agent: Enhancing efficiency with dynamic frame sampling and feedback-driven reasoning. *arXiv preprint arXiv:2410.20252*, 2024.

Susmit Jha, Vasumathi Raman, Dorsa Sadigh, and Sanjit A Seshia. Safe autonomy under perception uncertainty using chance-constrained temporal logic. *Journal of Automated Reasoning*, 60:43–62, 2018.

Fan Jiang, Junsong Yuan, Sotirios A. Tsaftaris, and Aggelos K. Katsaggelos. Anomalous video event detection using spatiotemporal context. *Comput. Vis. Image Underst.*, 115(3):323–333, 2011. doi: 10.1016/J.CVIU.2010.10.008. URL https://doi.org/10.1016/j.cviu.2010.10.008.

Hadas Kress-Gazit, Georgios E Fainekos, and George J Pappas. Temporal-logic-based reactive mission and motion planning. *IEEE transactions on robotics*, 25(6):1370–1381, 2009.

Aliaksandr Kroshchanka, Vladimir Golovko, Egor Mikhno, Mikhail Kovalev, Vadim Zahariev, and Aleksandr Zagorskij. A neural-symbolic approach to computer vision. In *International Conference on Open Semantic Technologies for Intelligent Systems*, pages 282–309. Springer, 2021.

Pika Labs. Pika ai: Free video generator with scene ingredients, 2024. URL https://pikartai.com. Pika 2.1 documentation.

Nanjun Li, Faliang Chang, and Chunsheng Liu. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. *Neurocomputing*, 490:482–494, 2022.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Effrosyni Mavroudi, Benjamín Béjar Haro, and René Vidal. Representation learning on visual-symbolic graphs for video understanding. In *European Conference on Computer Vision*, volume 12374 of *Lecture Notes in Computer Science*, pages 71–90. Springer, 2020.

Gérard G. Medioni, Isaac Cohen, François Brémond, Somboon Hongeng, and Ramakant Nevatia. Event detection and analysis from video streams. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(8):873–889, 2001.

Noushin Mehdipour, Matthias Althoff, Radboud Duintjer Tebbens, and Calin Belta. Formal methods to comply with rules of the road in autonomous driving: State of the art and grand challenges. *Automatica*, 152:110692, 2023.

Metricom. 8 must-follow social listening trends for 2025, January 2025. URL https://www.metricom.io/blog/social-listening-trends-2025.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022.

National Highway Traffic Safety Administration. Federal motor vehicle safety standards; automatic emergency braking systems for light vehicles. Technical Report Docket No. NHTSA-2023-0021, U.S. Department of Transportation, April 2024. URL https://www.nhtsa.gov/sites/nhtsa.gov/files/2024-04/final-rule-automatic-emergency-braking-systems-light-vehicles_web-version.pdf.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. Video generation models as world simulators, 2024. URL https://openai.com/sora/. Sora technical report.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023.

RapidAPI. Rapidapi hub. https://rapidapi.com. Accessed: 2025-03-02.

Runway Research. Introducing gen-3 alpha: A new frontier for video generation, 2024. URL https://runwayml.com/research/introducing-gen-3-alpha. Runway Gen-3 technical report.

Soumalya Sarkar, Kin Gwn Lore, and Soumik Sarkar. Early detection of combustion instability by neural-symbolic analysis on hi-speed video. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches co-located with the 29th Annual Conference on Neural Information Processing Systems*, volume 1583 of *CEUR Workshop Proceedings*, Montreal, Canada, 2015. CEUR-WS.org.

SP Sharan, Minkyu Choi, Sahil Shah, Harsh Goel, Mohammad Omama, and Sandeep Chinchali. Neuro-symbolic evaluation of text-to-video models using formalf verification. *arXiv preprint arXiv:2411.16718*, 2024.

Yasser Shoukry, Pierluigi Nuzzo, Ayca Balkan, Indranil Saha, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, George J Pappas, and Paulo Tabuada. Linear temporal logic motion planning for teams of underactuated robots using satisfiability modulo convex programming. In *2017 IEEE 56th annual conference on decision and control (CDC)*, pages 1132–1137. IEEE, 2017.

Richard H. Shultz and Richard D. Clarke. Big data at war: Special operations forces, project maven, and twenty-first-century warfare, August 2020.

Achint Soni, Sreyas Venkataraman, Abhranil Chandra, Sebastian Fischmeister, Percy Liang, Bo Dai, and Sherry Yang. Videoagent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2024.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *IEEE/CVF International Conference on Computer Vision*, pages 5551–5560. IEEE, 2019.

Twilio Inc. Twilio: Cloud communications platform, 2025. URL https://www.twilio.com. Accessed: 2025-03-01.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 699–714, 2024a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Discovqa: Temporal distortion-content transformers for video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4840–4854, 2023a.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023b.

Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *IEEE/CVF International Conference on Computer Vision*, pages 4591–4600. IEEE, 2019.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A discriminative CNN video representation for event detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1807, Boston, MA, USA, 2015. IEEE Computer Society.

Yunhao Yang, Jean-Raphaël Gaglione, Sandeep Chinchali, and Ufuk Topcu. Specification-driven video search via foundation models and formal verification. *arXiv preprint arXiv:2309.10171*, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 1039–1050, 2018.

Dongran Yu, Bo Yang, Qianhao Wei, Anchen Li, and Shirui Pan. A probabilistic graphical model based on neural-symbolic reasoning for visual relationship detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10599–10608, New Orleans, LA, USA, 2022. IEEE.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a. URL https://arxiv.org/abs/2306.02858.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023b.

Xiangtao Zheng, Yichao Zhang, Yunpeng Zheng, Fulin Luo, and Xiaoqiang Lu. Abnormal event detection by a weakly supervised temporal attention network. *CAAI Transactions on Intelligence Technology*, 7(3):419–431, 2022.