Novelty Assessment Report

Paper: VideoAgentTrek: Computer-Use Pretraining from Unlabeled Videos
PDF URL: https://openreview.net/pdf?id=xxYPqm1qWz
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Training computer-use agents requires massive amounts of GUI interaction data, but manually annotating action trajectories at scale is prohibitively expensive. We present VideoAgentTrek, a scalable pipeline that automatically mines training data from publicly available screen-recorded videos, eliminating the need for manual annotation. Our approach addresses a key challenge: raw videos contain implicit demonstrations but lack explicit action labels. To solve this, we develop Video2Action, an inverse dynamics module (IDM) with two components: (1) a video grounding model that detects and localizes GUI actions with precise temporal boundaries, and (2) an action-content recognizer that extracts structured parameters like click coordinates and typed text. Applied to 39,000 YouTube tutorial videos, our pipeline generates 1.52 million interaction steps. We leverage this data through continued pretraining followed by supervised fine-tuning. On OSWorld-Verified, our approach improves task success rates from 9.3% (SFT-only baseline) to 15.8%, a 70% relative improvement. On AgentNetBench, step accuracy increases from 64.1% to 69.3%. Our results demonstrate that passive internet videos can be transformed into high-quality supervision for computer-use agents, providing a scalable alternative to expensive manual annotation.

Core Task Landscape

This paper addresses: **Extracting GUI Action Trajectories from Unlabeled Screen-Recorded Videos**

A total of **30 papers** were analyzed and organized into a taxonomy with **12 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Automated Action Extraction and Recognition**
- **Large-Scale Data Mining and Agent Training**
- **Workflow and Behavioral Pattern Mining**
- **Human-Centered Analysis and Evaluation**
- **Programming by Demonstration Systems**

Complete Taxonomy Tree

- Extracting GUI Action Trajectories from Unlabeled Screen-Recorded Videos Survey Taxonomy
- Automated Action Extraction and Recognition
  - Inverse Dynamics and Action Parameter Extraction ★ (3 papers)
  - [0] VideoAgentTrek: Computer-Use Pretraining from Unlabeled Videos (Anon et al., 2026) View paper
  - [2] Learn to automate GUI tasks from demonstration (Intharah Thanapong, 2018) View paper
  - [17] GUI-Shift: Enhancing VLM-Based GUI Agents through Self-supervised Reinforcement Learning (Zhang, 2025) View paper
  - Video-to-Scenario Translation (2 papers)
  - [11] Translating Video Recordings of Complex Mobile App UI Gestures into Replayable Scenarios (Bernal-Cardenas, 2022) View paper
  - [12] V2S: A Tool for Translating Video Recordings of Mobile App Usages into Replayable Scenarios (Havranek, 2021) View paper
  - GUI Element Detection and Change Recognition (4 papers)
  - [4] Understanding screen relationships from screenshots of smartphone applications (Shirin Feiz, 2022) View paper
  - [25] Toward Mining Visual Log of Software (Pham Hung, 2016) View paper
  - [27] A Vision on Mining Visual Logs of Software (Hung Viet Pham, 2020) View paper
  - [29] Visual scrolling detection to enhance GUI agent training (Lee, n.d.) View paper
- Large-Scale Data Mining and Agent Training
  - Web Tutorial Mining for Agent Training (1 papers)
  - [8] TongUI: Internet-Scale Trajectories from Multimodal Web Tutorials for Generalized GUI Agents (Bofei Zhang, 2025) View paper
  - Self-Supervised and Reinforcement Learning for GUI Agents (2 papers)
  - [6] Assistgui: Task-oriented pc graphical user interface automation (Difei Gao, 2024) View paper
  - [18] GUI-explorer: Autonomous Exploration and Mining of Transition-aware Knowledge for GUI Agent (Chen Gong-wei, 2025) View paper
- Workflow and Behavioral Pattern Mining
  - Time-Series Interaction Data Extraction (3 papers)
  - [9] Extracting and analyzing time-series HCI data from screen-captured task videos (Lingfeng Bao, 2017) View paper
  - [13] Reverse engineering time-series interaction data from screen-captured videos (Lingfeng Bao, 2015) View paper
  - [14] BPMiner: mining developers' behavior patterns from screen-captured task videos (Jing Li, 2016) View paper
  - Process Mining and Workflow Optimization (3 papers)
  - [3] Process Mining IPTV Customer Eye Gaze Movement Using Discrete-Time Markov Chains (Zhi Chen, 2023) View paper

- [15] The Invisible Mentor: Inferring User Actions from Screen Recordings to Recommend Better Workflows (Yan, 2025) View paper
- [21] Integrating process mining and cognitive analysis to study EHR workflow (Stephanie K Furniss, 2017) View paper
- Interaction Mining for Large-Scale Applications (2 papers)
- [16] Simulating Web User Behavior Using LLM-Driven Browser Automation for Realistic IDS Dataset Generation (Ã□rdi, 2025) View paper
- [19] Democratizing interaction mining (Arsan, 2023) View paper

- Human-Centered Analysis and Evaluation
  - Usability Testing and Error Detection (3 papers)
  - [20] Usability testing of Avoiding Diabetes Thru Action Plan Targeting (ADAPT) decision support for integrating care-based counseling of pre-diabetes in an electronic â□¦ (D Chrimes, 2014) View paper
  - [22] Backtracking events as indicators of usability problems in creation-oriented applications (David Akers, 2012) View paper
  - [28] From error detection to behaviour observation: first results from screen capture analysis (Tort, 2009) View paper
  - Qualitative Behavioral Analysis (3 papers)
  - [1] A Processâ□□Oriented Approach to Analyze Analysts' Use of Visualizations: Revealing Insights into the What, When, and How (L. Zimmermann, 2025) View paper
  - [7] Bridging Quantitative and Qualitative Digital Experience Testing (Ranjitha Kumar, 2023) View paper
  - [24] Modeling user behavior and attention in search (Huang, 2013) View paper
  - Specialized Domain Applications (3 papers)
  - [10] Mechanism to capture learnerâ□□s interaction in VR-based learning environment: design and application (Rumana Pathan, 2020) View paper
  - [23] AMbER-Adaptive Instructional Systems as a Use Case for the Holistic Assessment Platform (Thomas E. F. Witte, 2023) View paper
  - [30] replication DOI: 11177/147337151EB21747 (D Gotz, n.d.) View paper
- Programming by Demonstration Systems (2 papers)
  - [5] Watch what I do: programming by demonstration (Allen Cypher, 1993) View paper
  - [26] Gesture morpher: video-based retargeting of multi-touch interactions (Ramik Sadana, 2016) View paper

Narrative

Core task: Extracting GUI action trajectories from unlabeled screen-recorded videos. This field addresses the challenge of automatically recovering structured interaction sequences—clicks, scrolls, text entries, and navigation steps—from raw video recordings of user sessions. The taxonomy reveals five main branches that reflect different motivations and technical emphases. Automated Action Extraction and Recognition focuses on the core computer vision and inverse dynamics problems: identifying what actions occurred and inferring their parameters from pixel-level observations, as exemplified by VideoAgentTrek[0] and Learn Automate GUI[2]. Large-Scale Data Mining and Agent Training leverages extracted trajectories to build datasets for training autonomous agents or recommendation systems. Workflow and Behavioral Pattern Mining seeks higher-level process models and usage patterns from aggregated interaction logs. Human-Centered Analysis and Evaluation emphasizes usability testing, user experience assessment, and understanding how people actually navigate interfaces. Finally, Programming by Demonstration Systems, rooted in early work like Watch What I Do[5], aims to enable end-users to teach software new behaviors by example, turning recorded actions into reusable scripts or macros.

Across these branches, a recurring tension exists between fully automated vision-based extraction and methods that rely on instrumentation or partial annotations. Within Automated Action Extraction, VideoAgentTrek[0] and its neighbors Learn Automate GUI[2] and GUI Shift[17] all tackle inverse dynamics—recovering action semantics from visual evidence alone—but differ in whether they assume access to underlying UI metadata or must work purely from pixels. VideoAgentTrek[0] emphasizes scalable, annotation-free extraction from diverse screen recordings, positioning itself closer to vision-centric approaches that handle arbitrary applications without prior instrumentation. In contrast, GUI Shift[17] explores domain adaptation when interface layouts change, and Learn Automate GUI[2] integrates demonstration learning with automation goals. Meanwhile, branches like Workflow Mining (e.g., BPMiner[14]) and Human-Centered Evaluation (e.g., IPTV Eye Gaze[3]) often assume richer input signals—event logs or gaze data—to study aggregate patterns or user experience, rather than solving the low-level action recognition problem. This landscape highlights an ongoing challenge: balancing the generality and scalability of pure video analysis against the precision afforded by instrumented or hybrid approaches.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Learn to automate GUI tasks from demonstration

Authors: Intharah Thanapong, Thanapong Intharah | **Year/Venue**: 2018 | **URL**: View paper

Abstract

This thesis explores and extends Computer Vision applications in the context of Graphical User Interface (GUI) environments to address the challenges of Programming by Demonstration (PbD). The challenges are explored in PbD which could be addressed through innovations in Computer Vision, when GUIs are treated as an application domain, analogous to automotive or factory settings. Existing PbD systems were restricted by domain applications or special application interfaces. Although they use the t...

Relationship Analysis

Both papers belong to the Inverse Dynamics and Action Parameter Extraction category, focusing on inferring GUI actions from visual observations. VideoAgentTrek extracts action trajectories from unlabeled screen-recorded videos using a two-stage inverse dynamics module (action event detection and parameter extraction) to generate large-scale training data for computer-use agents, while the candidate paper addresses learning GUI task automation from user demonstrations through visual-based programming by demonstration, focusing on teaching systems to replicate demonstrated tasks. The key difference is that VideoAgentTrek targets scalable pretraining from passive web videos without manual annotation, whereas the candidate emphasizes interactive learning from direct user demonstrations with human-in-the-loop refinement.

---

2. GUI-Shift: Enhancing VLM-Based GUI Agents through Self-supervised Reinforcement Learning

Authors: Zhang, Li, Longxi Gao, Gao, Pengzhi, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Training effective Vision-Language Models (VLMs) for GUI agents typically depends on large-scale annotated datasets, whose collection is both labor-intensive and error-prone. We introduce K-step GUI Transition, a self-supervised inverse dynamics task in which VLMs learn GUI dynamics by predicting the initial action that causes a transition between two GUI states. This approach eliminates the need for natural language instructions and enables scalable dataset construction from existing GUI trajec...

Relationship Analysis

Both papers belong to the Inverse Dynamics and Action Parameter Extraction category, focusing on learning GUI action trajectories from visual observations without extensive manual annotation. While VideoAgentTrek extracts structured action parameters (click coordinates, typed text) from unlabeled screen-recorded videos using a two-stage inverse dynamics module (action event detection + parameterization), GUI-Shift employs a self-supervised K-step GUI Transition task that predicts initial actions causing state transitions, combined with reinforcement learning for optimization. The key difference is that VideoAgentTrek focuses on mining training data from passive tutorial videos through explicit action extraction, whereas GUI-Shift learns GUI dynamics through self-supervised RL on state transitions without requiring natural language instructions.

Contributions Analysis

Overall novelty summary. The paper introduces VideoAgentTrek, a pipeline for mining GUI training data from unlabeled YouTube videos, and Video2Action, an inverse dynamics module that detects actions and extracts parameters like click coordinates. It resides in the 'Inverse Dynamics and Action Parameter Extraction' leaf, which contains only three papers total. This leaf sits within the broader 'Automated Action Extraction and Recognition' branch, indicating a relatively sparse research direction focused specifically on inferring structured action parameters from visual observations without manual annotation.

The taxonomy reveals neighboring leaves addressing related but distinct problems: 'Video-to-Scenario Translation' (two papers) focuses on converting recordings into replayable sequences, while 'GUI Element Detection and Change Recognition' (four papers) emphasizes identifying UI components rather than action semantics. The sibling papers in the same leaf—Learn Automate GUI and GUI Shift—both tackle inverse dynamics but differ in scope: one integrates demonstration learning with automation, the other addresses domain adaptation under layout changes. VideoAgentTrek's emphasis on large-scale, annotation-free extraction from diverse web videos positions it at the intersection of automated extraction and the 'Web Tutorial Mining for Agent Training' leaf (one paper), which also leverages online tutorials but may assume different input modalities or annotation strategies.

Among 22 candidates examined, the contribution-level analysis shows mixed novelty signals. The VideoAgentTrek pipeline (10 candidates examined, 1 refutable) and the Video2Action module (2 candidates examined, 1 refutable) both face at least one overlapping prior work within the limited search scope. The two-stage training methodology (10 candidates examined, 2 refutable) shows the most substantial prior overlap. These statistics suggest that while the specific combination may be novel, individual components—inverse dynamics modeling, video grounding, and pretraining-then-finetuning—have precedents in the examined literature. The relatively small candidate pool (22 papers) and sparse taxonomy leaf (3 papers) indicate this assessment is based on a focused but not exhaustive search.

Given the limited search scope and the sparse taxonomy leaf, the work appears to occupy a less-crowded research direction within GUI action extraction. The contribution-level statistics indicate that some technical components have prior work among the 22 candidates examined, but the integration of large-scale web video mining with inverse dynamics for agent training may represent a novel synthesis. A more exhaustive search across adjacent fields—such as video understanding, robotics imitation learning, or broader GUI automation—would be needed to fully assess originality beyond the top-K semantic matches analyzed here.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: VideoAgentTrek pipeline for mining training data from unlabeled videos

Description: The authors introduce a scalable pipeline that converts publicly available screen-recorded tutorial videos into structured training data for computer-use agents without requiring manual annotation. This approach addresses the data bottleneck in training GUI agents by leveraging abundant internet videos.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. UI-Hawk: Unleashing the Screen Stream Understanding for GUI Agents

URL: View paper

Brief Assessment

UI-Hawk[49] focuses on processing screen streams with a history-aware visual encoder for GUI navigation, not on mining training data from unlabeled videos. The candidate does not address automated trajectory extraction from tutorial videos.

---

2. ScreenAgent: A Vision Language Model-driven Computer Control Agent

URL: View paper

Brief Assessment

ScreenAgent[43] focuses on constructing a dataset through human annotation and GPT-4V correction for computer control tasks, not on automatically mining training data from unlabeled screen-recorded videos. The paper states 'we invoke gpt-4v to generate an original response, and annotators correct this response as the golden labeled response' (Figure 3 caption), indicating a supervised annotation process rather than unsupervised video mining.

---

3. UICVD: A Computer Vision UI Dataset for Training RPA Agents

URL: View paper

Brief Assessment

UICVD[46] focuses on static screen captures from enterprise applications for UI component recognition in RPA, not on mining training data from unlabeled screen-recorded videos or temporal action extraction from video sequences.

---

4. UI-Genie: A Self-Improving Approach for Iteratively Boosting MLLM-based Mobile GUI Agents

URL: View paper

Brief Assessment

UI Genie[44] focuses on self-improving GUI agents through reward models and synthetic trajectory generation, not on mining training data from unlabeled screen-recorded videos. The candidate's data generation uses 'rule-based verification, controlled trajectory corruption, and hard negative mining' rather than extracting actions from raw tutorial videos.

---

5. GUI-explorer: Autonomous Exploration and Mining of Transition-aware Knowledge for GUI Agent

URL: View paper

Brief Assessment

GUI Explorer[18] focuses on autonomous exploration of mobile GUI environments and mining transition-aware knowledge from real-time interactions, not on extracting training data from pre-recorded unlabeled videos. The candidate's exploration generates new trajectories through active interaction, whereas the original paper processes existing screen-recorded tutorial videos.

---

6. TongUI: Building Generalized GUI Agents by Learning from Multimodal Web Tutorials

URL: View paper

Brief Assessment

TongUI Multimodal[47] focuses on processing multimodal web tutorials (videos and articles) from platforms like YouTube and WikiHow, using ASR and LLMs to extract structured trajectories. While both approaches mine data from online videos, TongUI's pipeline differs technically by combining article and video sources, using ASR for textual extraction, and employing zero-shot GUI agents for trajectory generation, rather than VideoAgentTrek's inverse dynamics module (video grounding + action-content recognizer) for direct action extraction from raw videos.

---

7. OmniParser for Pure Vision Based GUI Agent

URL: View paper

Brief Assessment

OmniParser[42] focuses on parsing user interface screenshots into structured elements to enhance action grounding, not on mining training data from unlabeled videos. The candidate addresses screen parsing and icon detection, while the original contribution concerns automated trajectory extraction from tutorial videos.

---

8. Harnessing webpage uis for text-rich visual understanding

URL: View paper

Brief Assessment

Webpage UIs[48] focuses on synthesizing multimodal instructions from webpage accessibility trees using text-based LLMs, not on mining training data from unlabeled screen-recorded videos. The candidate's approach processes static webpage structures rather than temporal video content with action sequences.

---

9. UI-E2I-Synth: Advancing GUI Grounding with Large-Scale Instruction Synthesis

URL: View paper

Brief Assessment

UI-E2I-Synth[45] focuses on GUI instruction grounding through synthetic data generation using GPT-4o, not on mining training data from unlabeled screen-recorded videos. The candidate addresses a different technical problem (mapping instructions to GUI elements) using a different approach (LLM-based synthesis rather than inverse dynamics from videos).

---

10. Watch and Learn: Learning to Use Computers from Online Videos

URL: View paper

Prior Art Analysis

Watch and Learn[50] demonstrates that similar prior work exists for converting unlabeled screen-recorded videos into training data for computer-use agents. Both papers address the same fundamental problem: transforming publicly available tutorial videos into structured action trajectories without manual annotation. Watch and Learn[50] explicitly frames this as an inverse dynamics problem and reports generating over 53k trajectories from internet videos, predating the original paper's submission. The candidate paper's approach of 'predicting user actions from consecutive screen states' and converting 'readily available internet videos of human computer use into executable ui trajectories at scale' directly parallels the original paper's VideoAgentTrek pipeline that 'automatically mines training data from publicly available screen-recorded videos' using inverse dynamics.

Evidence

Evidence 1 - **Rationale**: Both papers present frameworks that convert publicly available videos into training data at scale without manual annotation, demonstrating that Watch and Learn[50] addresses the identical problem space. - **Original**: we present videoagenttrek, a scalable pipeline that automatically mines training data from publicly available screen-recorded videos, eliminating the need for manual annotation. - **Candidate**: we present watch&learn (w&l), a framework that converts readily available internet videos of human computer use into executable ui trajectories at scale.

Evidence 2 - **Rationale**: Both papers explicitly frame the solution as an inverse dynamics problem to extract action labels from unlabeled videos, showing Watch and Learn[50] employed the same core technical approach. - **Original**: our approach addresses a key challenge: raw videos contain implicit demonstrations but lack explicit action labels. to solve this, we develop video2action, an inverse dynamics module (idm) - **Candidate**: instead of directly generating actions or relying on handcrafted heuristics, we cast trajectory annotation as an inverse dynamics problem that predicts user actions from consecutive screen states

Evidence 3 - **Rationale**: Both papers identify the same data bottleneck problem and propose converting passive video recordings into structured training data as the solution. - **Original**: we can learn to automatically extract structured action trajectories from raw videos by training specialized models to detectwhenactions occur and infer whattheir parameters are, effectively converting passive recordings into active training data. - **Candidate**: computer-using agents (cuas) must plan task workflows across diverse and evolving applications, yet progress is limited by the lack of large-scale, high-quality training data. existing datasets are narrow, static, and costly to annotate

Evidence 4 - **Rationale**: Watch and Learn[50] demonstrates a complete implementation that generates tens of thousands of trajectories from videos, proving the approach was already realized before the original submission. - **Original**: teaching machines to use computers like humans do (clicking buttons, typing text, navigating interfaces) represents a fundamental challenge in ai. while recent advances in vision-language models have made computer-use agents increasingly feasible, their development remains bottlenecked by data avail... - **Candidate**: Through a task-aware retrieval and labeling pipeline, w&l yields over 53k high-quality trajectories that enhance cuas both as in-context exemplars and as supervised training data.

Contribution 2: Video2Action inverse dynamics module

Description: The authors develop an inverse dynamics module that recovers action supervision from raw videos through two stages: detecting GUI action events with millisecond-precision temporal localization, and extracting structured action parameters such as click coordinates and typed text to produce complete training trajectories.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos

URL: View paper

Prior Art Analysis

Video PreTraining[31] demonstrates that inverse dynamics models can be trained to recover action labels from unlabeled videos, predicting actions from observations without ground-truth labels. The paper explicitly describes training an IDM that 'seeks to minimize the negative log-likelihood of an action at timestep t given a trajectory of t observations' and uses it to label large-scale video data. This approach of using inverse dynamics to extract action supervision from videos predates the original paper's Video2Action module, which similarly detects GUI actions and extracts parameters from raw videos.

Evidence

Evidence 1 - **Rationale**: Both papers describe inverse dynamics modules that recover action labels from video observations. Video PreTraining[31] explicitly trains an IDM to predict actions from observations, establishing prior work on using inverse dynamics for action extraction from videos. - **Original**: we develop video2action, an inverse dynamics module (idm) with two components: (1) a video grounding model that detects and localizes gui actions with precise temporal boundaries, and (2) an action-content recognizer that extracts structured parameters like click coordinates and typed text. - **Candidate**: inverse dynamics models (idm) vpt, illustrated in figure 2, requires we first collect a small amount of labeled contractor data with which to train an inverse dynamics model pidm (at| o1...t), which seeks to minimize the negative log-likelihood of an action at timestep tgiven a trajectory of t observa...

Evidence 2 - **Rationale**: Both papers identify the same core challenge (videos lack action labels) and propose the same solution approach (inverse dynamics models). Video PreTraining[31] explicitly discusses how IDMs can use both past and future frames to infer actions, similar to the original paper's approach. - **Original**: our approach addresses a key challenge: raw videos contain implicit demonstrations but lack explicit action labels. to solve this, we develop video2action, an inverse dynamics module (idm) - **Candidate**: in contrast to an imitation learning policy, the idm can be non-causal, meaning its prediction for at can be a function of both past and future events, i.e. ot′>t. compared to the behavioral cloning objective of modeling the distribution of human intent given past frames only, we hypothesize that in...

---

2. ARDuP: Active Region Video Diffusion for Universal Policies

URL: View paper

Brief Assessment

ARDuP[32] focuses on inverse dynamics for robotic manipulation tasks using motion cues and latent representations for action decoding, not on GUI action detection with millisecond-precision temporal localization and structured parameter extraction (click coordinates, typed text) from screen recordings.

Contribution 3: Two-stage training methodology combining video pretraining and supervised fine-tuning

Description: The authors propose a training methodology that first performs continued pretraining on large-scale automatically mined video trajectories to learn fundamental GUI interaction patterns, then applies supervised fine-tuning on curated data to sharpen task-specific performance, demonstrating substantial improvements over fine-tuning alone.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Knowledge-guided pre-training and fine-tuning: Video representation learning for action recognition

URL: View paper

Brief Assessment

Knowledge Guided Pretraining[41] focuses on video representation learning for action recognition in videos, not on training computer-use agents through GUI interaction trajectories. The two-stage approach in the candidate addresses video understanding tasks, while the original paper's methodology specifically targets automated GUI agent training from screen recordings.

---

2. Reinforcement learning with action-free pre-training from videos

URL: View paper

Prior Art Analysis

Action Free Pretraining[33] demonstrates a two-stage training methodology that predates the ORIGINAL paper's approach. The candidate paper explicitly describes pre-training an action-free latent video prediction model on videos, followed by fine-tuning with action-conditional models on downstream tasks. This framework consists of '(i) pre-train an action-free latent video prediction model, and then utilize the pre-trained representations for efficiently learning action-conditional world models on unseen environments.' The ORIGINAL paper's contribution of 'continued pretraining on large-scale automatically mined video trajectories to learn fundamental GUI interaction patterns, then applies supervised fine-tuning on curated data' follows the same two-stage paradigm of video pretraining followed by supervised fine-tuning that Action Free Pretraining[33] established.

Evidence

Evidence 1 - **Rationale**: Both papers describe a two-stage training methodology where video pretraining is followed by fine-tuning. The candidate explicitly establishes this framework structure before the original paper. - **Original**: we validate videoagenttrekwith a two-stage training recipe: continued pretraining on the mined trajectories followed by supervised fine-tuning on a curated dataset. this combination leverages the broad coverage from videos to learn fundamental gui interaction patterns, while supervised fine-tuning s... - **Candidate**: our framework consists of two phases: we pre-train an action-free latent video prediction model, and then utilize the pre-trained representations for efficiently learning actionconditional world models on unseen environments.

Evidence 2 - **Rationale**: Both papers describe a stage 1 pretraining phase on video data followed by a stage 2 fine-tuning phase that incorporates task-specific supervision, demonstrating the same two-stage methodology. - **Original**: stage 1 training.we train for one epoch over26btokens drawn from the videoagenttrek trajectories, augmented with a small number of gui grounding pairs. stage 2 training.we continue training for8btokens on a curated set of clean, human-annotated trajectories. - **Candidate**: once we pre-train the action-free prediction model, we finetune it into an action-conditional prediction model that can be used for solving various visual control tasks. since actions and rewards, which provide more information about target tasks, are available during fine-tuning, it motivates incorpo...

---

3. Towards learning a generic agent for vision-and-language navigation via pre-training

URL: View paper

Brief Assessment

Vision Language Navigation[38] focuses on pre-training vision-language encoders for navigation tasks using image-text-action triplets, not video pretraining from unlabeled screen recordings. The two-stage approach in [38] involves pre-training on navigation data followed by task-specific fine-tuning, which differs fundamentally from the original paper's video mining pipeline and continued pretraining on automatically extracted GUI trajectories.

---

4. Inter-Slice Super-Resolution of Magnetic Resonance Images by Pre-Training and Self-Supervised Fine-Tuning

URL: View paper

Brief Assessment

Inter Slice MRI[36] focuses on medical image super-resolution using video pretraining for MRI slice interpolation, not GUI agent training with action trajectories. The domains and tasks are fundamentally different.

---

5. AdaWorld: Learning Adaptable World Models with Latent Actions

URL: View paper

Brief Assessment

AdaWorld[34] focuses on world model pretraining with latent actions for visual planning, not on computer-use agent training with video pretraining followed by supervised fine-tuning for GUI interaction tasks.

---

6. Two-Stage Learning Approach for Semantic-Aware Task Scheduling in Container-Based Clouds

URL: View paper

Brief Assessment

Semantic Task Scheduling[40] addresses container-based cloud task scheduling using a two-stage pipeline with image-text pretraining followed by RL fine-tuning, which is fundamentally different from the original paper's video-based GUI agent training approach.

---

7. Fine-Tuning Video-Text Contrastive Model for Primate Behavior Retrieval from Unlabeled Raw Videos

URL: View paper

Brief Assessment

Primate Behavior Retrieval[39] focuses on fine-tuning video-text contrastive models for primate behavior retrieval using a data cleaning pipeline followed by LoRA-based fine-tuning. This differs fundamentally from the original paper's approach of continued pretraining on automatically mined GUI interaction trajectories followed by supervised fine-tuning for computer-use agents. The domains (animal behavior vs. GUI automation), data sources (wildlife videos vs. screen recordings), and technical implementations are entirely distinct.

---

8. Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos

URL: View paper

Prior Art Analysis

Video PreTraining[31] explicitly proposes and demonstrates a two-stage training methodology where models are first pretrained on large-scale video data labeled by an IDM, then fine-tuned on smaller, curated datasets. The paper shows that this approach substantially outperforms training from scratch and demonstrates improvements across multiple tasks. This establishes clear prior work on the two-stage video pretraining followed by supervised fine-tuning methodology.

Evidence

Evidence 1 - **Rationale**: Video PreTraining[31] describes training a foundation model on broad video data followed by fine-tuning to narrower, task-specific datasets, establishing the same two-stage methodology claimed as novel by the original paper. - **Original**: we validate videoagenttrekwith a two-stage training recipe: continued pretraining on the mined trajectories followed by supervised fine-tuning on a curated dataset. this combination leverages the broad coverage from videos to learn fundamental gui interaction patterns, while supervised fine-tuning s... - **Candidate**: foundation models are designed to have a broad behavior profile and be generally capable across a wide variety of tasks. to incorporate new knowledge or allow them to specialize on a narrower task distribution, it is common practice to fine-tune these models to smaller, more specific datasets. 1 the vp...

Evidence 2 - **Rationale**: Both papers demonstrate substantial improvements from the two-stage approach compared to baselines. Video PreTraining[31] shows quantitative improvements from fine-tuning after video pretraining, validating the methodology's effectiveness. - **Original**: experiments demonstrate that our approach achieves 15.8% task success on osworld-verified compared to 9.3% for sft-only baselines (70% relative improvement), and improves step accuracy on agentnetbench from 64.1% to 69.3%, validating that passive internet videos can provide effective supervision at ... - **Candidate**: fine-tuning to earlygame_keyword results in a large boost compared to the zero-shot foundation model: 2.5x more crafting tables, 6.1x more planks, 4.3x more logs, and 5.5x more crafting overall (fig. 5). however, when fine-tuning to this dataset we did not see any new behaviors emerge, only a refineme...

---

9. Bootstrapping World Models from Dynamics Models in Multimodal Foundation Models

URL: View paper

Brief Assessment

Bootstrapping World Models[37] focuses on bootstrapping world models from dynamics models using synthetic data and inference-time verification for action-centric image editing, not on video pretraining followed by supervised fine-tuning for GUI agents.

---

10. What Comes After Transformers? A Selective Survey Connecting Ideas in Deep LearningGPT

URL: View paper

Brief Assessment

After Transformers[35] is a survey paper discussing architectural alternatives to transformers and general deep learning patterns. It does not propose or evaluate any specific two-stage training methodology combining video pretraining with supervised fine-tuning for GUI agents or similar systems.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] VideoAgentTrek: Computer-Use Pretraining from Unlabeled Videos View paper
- [1] A Processâ□□Oriented Approach to Analyze Analysts' Use of Visualizations: Revealing Insights into the What, When, and How View paper
- [2] Learn to automate GUI tasks from demonstration View paper
- [3] Process Mining IPTV Customer Eye Gaze Movement Using Discrete-Time Markov Chains View paper
- [4] Understanding screen relationships from screenshots of smartphone applications View paper
- [5] Watch what I do: programming by demonstration View paper
- [6] Assistgui: Task-oriented pc graphical user interface automation View paper
- [7] Bridging Quantitative and Qualitative Digital Experience Testing View paper
- [8] TongUI: Internet-Scale Trajectories from Multimodal Web Tutorials for Generalized GUI Agents View paper
- [9] Extracting and analyzing time-series HCI data from screen-captured task videos View paper
- [10] Mechanism to capture learnerâ□□s interaction in VR-based learning environment: design and application View paper
- [11] Translating Video Recordings of Complex Mobile App UI Gestures into Replayable Scenarios View paper
- [12] V2S: A Tool for Translating Video Recordings of Mobile App Usages into Replayable Scenarios View paper
- [13] Reverse engineering time-series interaction data from screen-captured videos View paper
- [14] BPMiner: mining developers' behavior patterns from screen-captured task videos View paper
- [15] The Invisible Mentor: Inferring User Actions from Screen Recordings to Recommend Better Workflows View paper
- [16] Simulating Web User Behavior Using LLM-Driven Browser Automation for Realistic IDS Dataset Generation View paper

- [17] GUI-Shift: Enhancing VLM-Based GUI Agents through Self-supervised Reinforcement Learning View paper
- [18] GUI-explorer: Autonomous Exploration and Mining of Transition-aware Knowledge for GUI Agent View paper
- [19] Democratizing interaction mining View paper
- [20] Usability testing of Avoiding Diabetes Thru Action Plan Targeting (ADAPT) decision support for integrating care-based counseling of pre-diabetes in an electronic â View paper
- [21] Integrating process mining and cognitive analysis to study EHR workflow View paper
- [22] Backtracking events as indicators of usability problems in creation-oriented applications View paper
- [23] AMbER-Adaptive Instructional Systems as a Use Case for the Holistic Assessment Platform View paper
- [24] Modeling user behavior and attention in search View paper
- [25] Toward Mining Visual Log of Software View paper
- [26] Gesture morpher: video-based retargeting of multi-touch interactions View paper
- [27] A Vision on Mining Visual Logs of Software View paper
- [28] From error detection to behaviour observation: first results from screen capture analysis View paper
- [29] Visual scrolling detection to enhance GUI agent training View paper
- [30] replication DOI: 11177/147337151EB21747 View paper
- [31] Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos View paper
- [32] ARDuP: Active Region Video Diffusion for Universal Policies View paper
- [33] Reinforcement learning with action-free pre-training from videos View paper
- [34] AdaWorld: Learning Adaptable World Models with Latent Actions View paper
- [35] What Comes After Transformers? A Selective Survey Connecting Ideas in Deep LearningGPT View paper
- [36] Inter-Slice Super-Resolution of Magnetic Resonance Images by Pre-Training and Self-Supervised Fine-Tuning View paper
- [37] Bootstrapping World Models from Dynamics Models in Multimodal Foundation Models View paper
- [38] Towards learning a generic agent for vision-and-language navigation via pre-training View paper
- [39] Fine-Tuning Video-Text Contrastive Model for Primate Behavior Retrieval from Unlabeled Raw Videos View paper
- [40] Two-Stage Learning Approach for Semantic-Aware Task Scheduling in Container-Based Clouds View paper
- [41] Knowledge-guided pre-training and fine-tuning: Video representation learning for action recognition View paper
- [42] OmniParser for Pure Vision Based GUI Agent View paper
- [43] ScreenAgent: A Vision Language Model-driven Computer Control Agent View paper
- [44] UI-Genie: A Self-Improving Approach for Iteratively Boosting MLLM-based Mobile GUI Agents View paper
- [45] UI-E2I-Synth: Advancing GUI Grounding with Large-Scale Instruction Synthesis View paper
- [46] UICVD: A Computer Vision UI Dataset for Training RPA Agents View paper
- [47] TongUI: Building Generalized GUI Agents by Learning from Multimodal Web Tutorials View paper
- [48] Harnessing webpage uis for text-rich visual understanding View paper
- [49] UI-Hawk: Unleashing the Screen Stream Understanding for GUI Agents View paper
- [50] Watch and Learn: Learning to Use Computers from Online Videos View paper