

AI-Resistant Video CAPTCHA System

Ananya R. Bhat¹, Srushti Ravindra Chigadani², Yathin G. Kummar^{3,*}, Rashmi T.V.⁴

Abstract

CAPTCHA, which stands for Completely Automated Public Turing test to tell Computers and Humans Apart, is a tool that separates human users from automated bots. It is commonly used in web applications to stop bulk registrations, spam submissions, credential stuffing, and misuse of online services. Typically, CAPTCHAs involve recognizing distorted text, selecting images, or transcribing audio. These tasks are easy for people but challenging for automated programs. However, recent advances in artificial intelligence have made these methods less effective. Machine learning models and multimodal AI systems can now solve text, image, and audio CAPTCHAs with nearly human-like or even better accuracy. This poses serious risks, such as creating fake accounts, ticket scalping, and large-scale automated misuse of websites. As a result, researchers are looking into new methods, like video-based CAPTCHAs. These use short dynamic clips that introduce motion or timing challenges, relying on temporal reasoning. Such tasks are straightforward for humans but still difficult for AI, which often has trouble tracking changes across frames or recognizing complex motion boundaries. This survey and analysis show that static CAPTCHAs are no longer enough in the age of AI. Video-based CAPTCHAs, which include replay protection through hashing and timestamping along with adversarial testing to counter new attacks, provide a more practical and sustainable solution for modern web security.

Keywords: Adversarial testing, AI-resistant security, bot detection, temporal reasoning, video CAPTCHA, web application security.

INTRODUCTION

With the rise of artificial intelligence, traditional CAPTCHA systems, such as text puzzles, image selection, and audio challenges, have become less reliable. Bots can now solve these problems with high accuracy. This leads to serious problems, such as fake account creation, misuse of online services, and unfair market practices. Video-based CAPTCHAs offer a better solution to this problem. They show short moving clips in which users perform tasks related to motion or timing. These tasks are natural for humans but are difficult for automated programs. The addition of cryptographic hashing, timestamps, and continuous testing strengthens and secures the system. This approach fits real-world applications better than the older CAPTCHA methods.

*Author for Correspondence

Yathin G Kummar

E-mail: yathinkummar@gmail.com

¹⁻³Student, Department of ISE, BNM Institute of Technology
Bengaluru, India

⁴Assistant Professor, Department of ISE, BNM Institute of
Technology Bengaluru, India

Received Date: February 03, 2026

Accepted Date: March 12, 2026

Published Date: April 10, 2026

Citation: Ananya R Bhat, Srushti Ravindra Chigadani, Yathin G Kummar, Rashmi T V. AI-Resistant Video CAPTCHA System. Journal of Web Engineering & Technology. 2026; 13(2): 1–7p.

In text-based CAPTCHAs, users read and type distorted letters and numbers. Image-based CAPTCHAs usually involve selecting the correct object, animal, or item from a group of images. Audio CAPTCHAs, often used by visually impaired users, require listening to unclear or noisy speech and typing what is heard. A newer version, Google's reCAPTCHA, simplifies this process by allowing users to tick a box or identify objects in images while monitoring their actions on the website. Although these methods have been effective in the past, modern AI and machine learning systems can now solve them accurately.

This situation calls for stronger and more effective CAPTCHA solutions to prevent such attacks.

The spread of automated bots is a significant threat to Internet platforms. This affects e-commerce, ticketing, banking, and social media applications. Automated agents are now widely used for large-scale fake registrations, credential stuffing, denial-of-service attacks, and manipulation of opinions. These activities threaten both user security and trust. The most recognized defense against these threats is the Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA). Since its introduction over 20 years ago, CAPTCHA has changed from distorted text identification tasks to challenges that include images and voice. Although these methods were initially effective, advances in machine learning have quickly weakened their security.

Convolutional neural networks can now solve text-based CAPTCHAs with impressive accuracy, often outperforming humans. Deep learning attacks achieve success rates between 93% and 100% on common CAPTCHA systems, whereas human success rates are much lower. Even more alarming, commercial platforms such as Google reCAPTCHA and CAPTCHA have also been bypassed using adversarial techniques. This raises concerns regarding the security of even the most advanced systems.

Another important issue is the usability. As CAPTCHA designers add more distortion to combat automated solvers, the difficulty for human users significantly increases. Recent evaluations have shown that the average human success rates, once above 85%, have now dropped below 70% for several common schemes [1]. The additional cognitive and visual strain results in frustration and longer completion times, directly harming the user experience. Audio CAPTCHAs, originally intended to support accessibility, have also proven vulnerable to modern speech recognition systems, further reducing their practical value [2].

To tackle these challenges, researchers are looking into video-based CAPTCHAs, which use temporal information such as object motion and sequence. Unlike static images, video tasks utilize temporal reasoning, a skill in which humans still outshine AI systems. For example, Nayak and Saini developed a system that uses short video clips alongside Visual Question Answering techniques. It asks users to identify object interactions and colors [3]. Another method, called BounTCHa, employs motion boundary recognition to distinguish between human and machine responses [4]. Additional research has considered combining video challenges with other methods, such as honeywords and location verification, for extra security [5].

However, current video CAPTCHA solutions have several drawbacks. Some require longer video clips of 10–20 s, which can lead to delays and demand more bandwidth [3]. Others primarily focus on object recognition rather than more complex motion reasoning, making them susceptible to vision-based models [4]. Additionally, adversarial testing against advanced multimodal systems such as VideoMAE and transformer-based vision-language models remains scarce. This leaves unresolved questions regarding their long-term effectiveness [6].

To address these gaps, this study introduces a new AI-resistant video CAPTCHA system. The system generates short (2–3 s) video clips that include motion-based and boundary-oriented cues, such as detecting crossings, sudden direction changes, or unusual motion patterns. To ensure secure deployment, each challenge is cryptographically hashed and timestamped to protect against replay and injection attacks [7].

Furthermore, the framework includes an adversarial-feedback module. State-of-the-art AI solvers were tested against the CAPTCHA generator, allowing the system to adapt as machine learning evolves [8]. The main contributions of this study are as follows:

1. Development of a video CAPTCHA framework using short dynamic clips with motion and temporal reasoning.
2. Integration of an adaptive adversarial testing module to assess and strengthen CAPTCHA design against new AI models.
3. Implementation of session binding and replay protection through hashing and timestamping.
4. Experimental evaluation shows that humans maintain success rates of 80–90% with low cognitive load, whereas AI solvers are notably less effective.

The aim is to create a multilayered authentication system that separates humans from automated systems using cognitive, interactive, and perceptual challenges. It should maintain high usability, with a human success rate greater than 90% and low bot acceptance under 0.01%. [8]

Limitations of Current ReCAPTCHA System

Adaptive Threats from Emerging AI Models

- Modern deep learning and multimodal AI systems, such as Vision Transformers and GPT-like models, can now solve image- and text-based reCAPTCHA with human-level or better accuracy. CAPTCHAs may decrease over time.
- This weakens the system's ability to differentiate between bots and real users.

Reduced User Accessibility and Usability

As reCAPTCHA introduces more complex challenges, genuine users, especially those with visual or hearing impairments, have a more difficult time solving them.

Reliance on Behavioral and Tracking Data

reCAPTCHA v3 mainly relies on user behavior, cookies, and browsing patterns for risk analyses.

METHODOLOGY

The development of an AI-resistant video-based CAPTCHA system aims to create a secure yet user-friendly verification mechanism that can effectively distinguish between humans and automated agents. The methodology adopted in this study involved a systematic process of analysis, design, implementation, and evaluation. It draws insights from prior research on CAPTCHA security, motion-based verification, and human-computer interaction while addressing the vulnerabilities exposed by recent advancements in artificial intelligence. Traditional text- and image-based CAPTCHAs, which once served as reliable security measures, are increasingly being bypassed by AI systems such as GPT-4 Vision and YOLO-based models that achieve near-perfect solve rates. To overcome these limitations, the proposed methodology focuses on integrating video-based, time-dependent reasoning challenges that require perceptual and temporal understanding—abilities in which humans still hold a distinct advantage over machines [1, 2].

The process begins with user credential verification, where passwords are compared with secure hashes stored in the database. A honeyword mechanism is used to detect unauthorized login attempts to strengthen this layer. The system stores a mix of real and fake passwords, triggering alerts or redirections to honeypots whenever a decoy password is entered [4]. This step ensures that brute-force and credential-stuffing attempts are detected early. Additionally, IP and geolocation tracking are employed to detect anomalies such as logins from unusual locations. In such cases, the CAPTCHA difficulty is automatically increased to mitigate potential automated access without affecting legitimate users [7]. Once the authentication pre-check is completed, a short dynamic video CAPTCHA is created. These clips, typically 3–5 s long, are stored in secure cloud storage, such as AWS S3 or Google Cloud Buckets. Each video contained motion-based activities, such as an object crossing a boundary, changing direction, or performing a specific action. The backend dynamically selects a video and retrieves its metadata, which is stored in MongoDB, along with unique identifiers

and timestamps. Inspired by designs such as vrCAPTCHA [9] and BounTCHA [3], this approach introduces motion-based cues and boundary changes that cannot be easily predicted or recognized by AI models trained only on static visual data. To enhance variability, AI-assisted prompt generation tools can be used so that both the video and its questions are automatically produced from the same contextual description, reducing manual effort while keeping challenges unique.

Once a challenge is selected, the system automatically generates a question based on the video's metadata. This is done using vision-language techniques similar to the Visual Question Answering (VQA) framework [10]. For example, a video showing a car moving across a frame might generate the question, "In which direction did the car move?" or "How many times did the car cross the screen?" These closed-ended questions prevent ambiguity and ensure that the task remains simple for humans. Users either select an answer or track an object using their cursor or touch input. The movement data from this interaction were analyzed using mean-shift clustering to measure smoothness and responsiveness, as shown in Figure 1. Human users tend to display natural and continuous cursor motion, whereas bots exhibit sudden or mechanical movement patterns, allowing accurate differentiation [8].

The user's response is then validated through a secure backend mechanism. Each challenge is uniquely identified by a cryptographically signed token that prevents tampering or reuse. The correct answers are stored as salted hashes and are never sent to the client, ensuring that they cannot be intercepted. Once a response is submitted, the backend verifies the token signature, challenge validity, and hash match while checking for replay or injection attempts. Pre-signed video URLs are also configured to expire within 30–60 s, which eliminates the risk of unauthorized replays or dataset harvesting. All challenge interactions were logged and timestamped for auditing and further analysis.

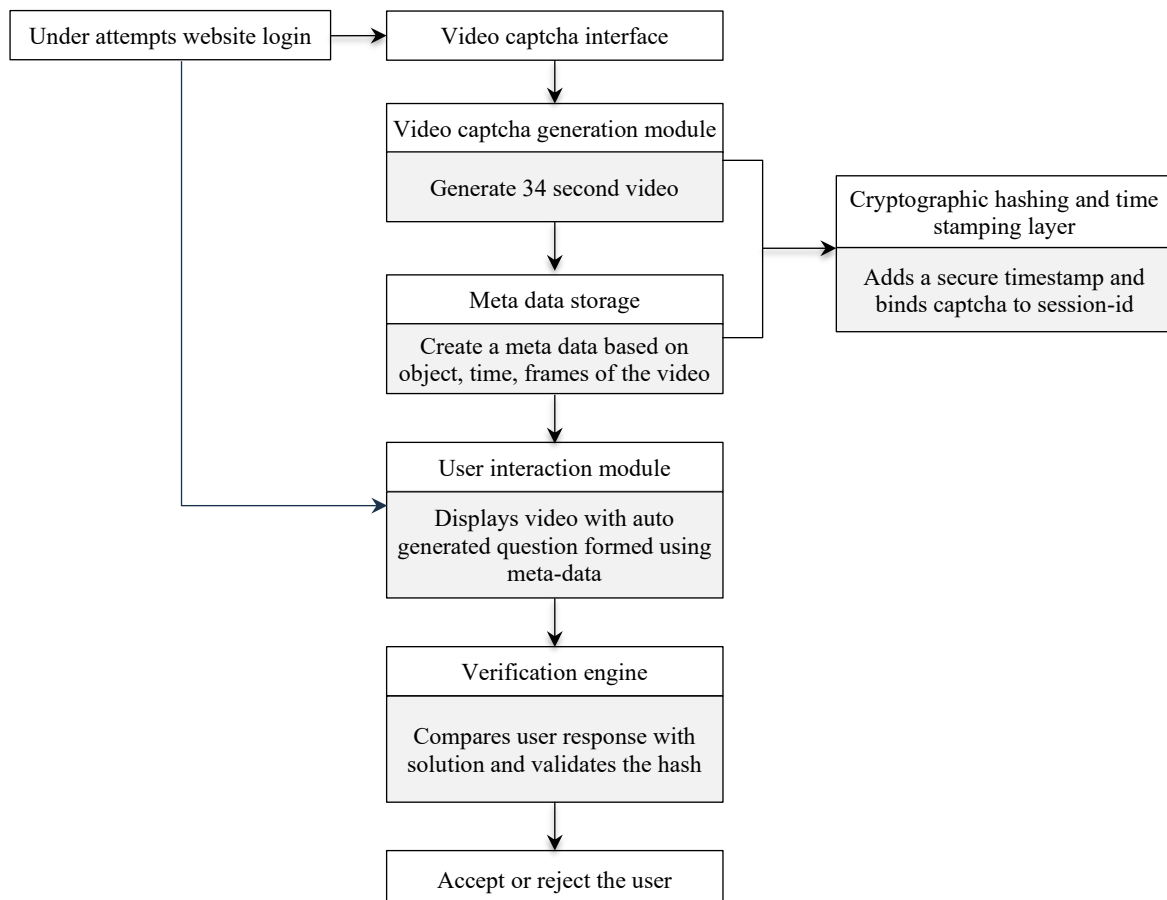


Figure 1. Proposed block architecture.

The system undergoes continuous evaluation through human testing and simulated AI-based attacks. Key performance indicators, such as solution accuracy, latency, and false acceptance rate, were measured. Studies have shown that humans can maintain success rates above 90%, whereas AI solvers remain under 0.1% when exposed to dynamic motion tasks [9, 5]. To ensure adaptability, the methodology includes adversarial testing, where the CAPTCHA is continuously exposed to AI solvers during the design phase, and the parameters are adjusted to maintain resistance. This cyclical improvement process ensures that the system remains robust against emerging models capable of multimodal understanding.

The implementation was carried out using a Python-based backend framework, such as Flask or Django, integrated with OpenCV for video frame analysis and PyTorch for AI-assisted feature extraction. MongoDB is used to store metadata, challenge information, and hashed responses, whereas cloud platforms such as AWS or GCP handle video storage and delivery using short-lived pre-signed URLs. The frontend, developed using React.js, retrieves the challenges from the backend through secure REST APIs and handles user interactions. Logging and monitoring were implemented using Redis and the ELK Stack for scalability and performance tracking.

Through this layered methodology, the system successfully combines perceptual, cognitive, and behavioral analyses into a single verification pipeline. Unlike static CAPTCHAs that rely solely on image recognition, this approach introduces dynamic motion reasoning and adversarial adaptability, thereby creating a resilient framework that evolves alongside AI advancements. The outcome is a secure, adaptive, and scalable CAPTCHA mechanism that maintains usability for legitimate users while significantly reducing the success rate of automated CAPTCHA solvers.

RESULTS AND DISCUSSION

The ongoing evolution of CAPTCHA technologies over the past two decades shows the continuous struggle between user-focused design and the growing abilities of automated solvers. Early methods, such as text distortion and audio transcription, offered significant barriers against bots [1]. However, rapid advancements in artificial intelligence have steadily weakened its effectiveness. With the rise of deep learning and multimodal large language models, many previously reliable CAPTCHA schemes are now consistently bypassed [2]. Consequently, researchers have suggested various improvements, including image-based challenges using diffusion models [8] and immersive virtual reality CAPTCHAs [4].

Table 1 summarizes the representative approaches from 2008 to 2025. It outlines the key techniques, reported advantages, and observed weaknesses. The trend clearly shows that while each system introduced valuable improvements, such as higher robustness, immersive experiences, or large-scale deployability, none has fully addressed the challenge of resisting modern adversarial attacks.

AI. For example, audio CAPTCHAs that once provided accessibility have been broken by speech recognition to block replay or injection attacks, along with adversarial testing to adapt the system against evolving AI threats. The proposed video-based CAPTCHA framework provides a sustainable pathway forward in strengthening web security.

Future Scope

Looking ahead, the proposed video-based CAPTCHA system presents several promising areas for further development. An important next step is to integrate it into real-world platforms, such as e-commerce sites, banking portals, and ticketing systems, to evaluate its effectiveness on a larger scale. Because threat levels differ in various contexts, the system can adapt by changing the difficulty of the video challenge based on user behavior and the chance of a bot attempt.

Table 1. Comparison of existing captcha system.

Year	Core idea/technique	Reported advantages	Observed gaps/weaknesses
2025	Image CAPTCHAs enhanced with diffusion models	Higher robustness than simple image CAPTCHAs	Still static; advanced vision models bypass diffusion noise
2024	Tests GPT-like LLMs on CAPTCHA challenges	Shows cross- modal evaluation	Demonstrates LLMs breaking text & image CAPTCHAs with ease
2023	Motion-based CAPTCHA in VR	Innovative, immersive	Needs costly hardware; poor scalability
2022	AI-based adversarial attack testing on CAPTCHAs	Expose design flaws early	Proves that static CAPTCHAs cannot survive AI evolution
2021	Traffic-rate and behavioral pattern checks	Works passively, scales with infra	Still bypassed with AI + replay
2020	Automatic Speech Recognition (ASR) on audio CAPTCHAs	Empirical attack study	Achieved ~90% solve rate
2016	Short ad videos, answer questions	Users enjoyed format	Easy to memorize or brute-force once dataset leaked
2013	OCR on moving text CAPTCHAs	Quantified attack precision/recall	Broke 9/13 videos; text-based is fragile
2013	Video segmentation + query engine	Large-scale system pipeline	Replay attacks and memorization not solved
2009–2008	Video tag-based CAPTCHAs; human vs bot eval	Balanced usability & security	Popular-tag bias, retrieval attacks

Another key area is mobile and edge deployment, which ensures that the solution works well even for users with low bandwidth or devices with limited resources. Additionally, combining temporal video tasks with subtle biometric cues, such as mouse movements or touch gestures, can further enhance security. Finally, creating a standardized framework to evaluate CAPTCHA against the latest AI models would help keep the system relevant as adversarial technologies continue to advance.

Challenges

Although the system shows great promise, its implementation presents challenges. The first concern is finding the right balance between usability and security issues. If CAPTCHA tasks are too complicated, they might frustrate real users, undermining the goal of accessibility. Simultaneously, creating and delivering dynamic videos adds computational demands, which could strain servers during high-traffic periods. Accessibility is also crucial, as users with different abilities, particularly those with visual or hearing impairments, will need alternative forms of challenge to ensure inclusivity. Furthermore, adversarial AI models are advancing rapidly, and improvements in temporal reasoning could eventually diminish the advantages that humans currently have. This requires the system to continuously evolve and adjust. Finally, issues related to data privacy and compliance must be carefully managed, as handling video content must ensure that no sensitive or personal information is accidentally exposed or stored in systems with nearly 90% accuracy by 2020 [3]. Recent studies in 2024 have demonstrated that GPT-like models can successfully solve multimodal CAPTCHAs [2]. Even the latest image-based approaches, improved with diffusion models in 2025, remain fundamentally static in nature. This makes them vulnerable to computer vision transformers and generative AI [8].

The analysis of these results points to a recurring limitation across existing systems: a lack of temporal reasoning and dynamic variability. Humans are naturally skilled at interpreting motion and recognizing temporal changes. In contrast, AI still struggles to track sequential visual cues with consistent accuracy [7].

This gap offers an opportunity to design video-based CAPTCHAs that rely on motion- and boundary-oriented challenges. This is achieved by further incorporating hashing and timestamping.

CONCLUSION

A survey of existing CAPTCHA research from 2008 to 2025 shows a clear trend: early mechanisms, such as distorted text, image selection, and audio transcription, provided effective defenses against automated programs, but these systems have struggled against the rapid progress of artificial intelligence. Studies have shown that modern speech recognition models can transcribe audio CAPTCHAs with nearly 90% accuracy. Large multimodal models can increasingly tackle text- and image-based challenges at the human level or even better performance. However, recent innovations such as diffusion-based image CAPTCHAs and immersive VR CAPTCHAs have limitations owing to their static nature and practical deployment issues. Comparative studies indicate that methods using fixed datasets, text overlays, or popular tag systems are vulnerable to memorization, dataset leakage, and automated retrieval attacks. Large-scale implementations, such as Google's video CAPTCHA patent, provide some architectural protection but do not guarantee resilience against modern vision and language models. These results reveal a lack of temporal reasoning and dynamic variability as major weaknesses in different generations of CAPTCHA systems. Clearly, static formats—whether text, images, or audio—can no longer act as a dependable defence. The reviewed research highlights the urgent need for CAPTCHA designs that go beyond single-frame recognition and include temporal, adaptive, and adversarial tested features. These insights, influenced by the limitations found in earlier studies, pave the way for future research on strong, AI-resistant solutions.

REFERENCES

1. Kopp M, Nikl M, Holena M. Breaking CAPTCHAs with Convolutional Neural Networks. InITAT 2017 Sep (pp. 93-99).
2. Guerar M, Verderame L, Migliardi M, Palmieri F, Merlo A. Gotta CAPTCHA'Em all: a survey of 20 Years of the human-or-computer Dilemma. ACM Computing Surveys (CSUR). 2021 Oct 7;54(9):1-33.
3. Yusuf MO, Srivastava D, Singh D, Rathor VS. Multiview deep learning-based attack to break text-CAPTCHAs. International journal of machine learning and cybernetics. 2023 Mar;14(3):959-72.
4. Aubry P, Devoivre J, Carron D, Fernandez S, Duda A, Korczyński M. Bypassing Audio reCAPTCHA with Automatic Speech Recognition Models. 2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) 2025 Jun 30. pp. 1-5.
5. Johri E, Dharod L, Joshi R, Kulkarni S, Kundle V. Video Captcha Proposition based on VQA, NLP, Deep Learning and Computer Vision. In2022 5th International Conference on Advances in Science and Technology (ICAST) 2022 Dec 2. pp. 196-200.
6. Lin L, Wang K, Abdallah M, Cai W. BounTCHA: A CAPTCHA Utilizing Boundary Identification in Guided Generative AI-extended Videos. arXiv preprint arXiv:2501.18565. 2025 Jan 30.
7. Patil N, Patil R. Achieving flatness: with video captcha, location tracking, selecting the honeywords. In2018 International Conference on Smart City and Emerging Technology (ICSCET) 2018 Jan 5. pp. 1-6.
8. Wu Z, Xue Y, Feng Y, Wang X, Song Y. MCA-bench: A multimodal benchmark for evaluating CAPTCHA robustness against VLM-based attacks. In Proceedings of the AAAI Conference on Artificial Intelligence 2026 Mar 14 (Vol. 40, No. 44, pp. 38039-38047).
9. Gurale M, Borate K, Sangitrao RR, Tiwari G. Enhanced Security Through Honeywords and Video CAPTCHA: An Innovative Authentication and Tracking Mechanism. In2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) 2025 Jan 9. pp. 1-6.
10. Moradi M, Moradi M, Palazzo S, Rundo F, Spampinato C. Image CAPTCHAs: When deep learning breaks the mold. IEEE Access. 2024 Aug 13;12:112211-31.