

# BounTCHA: A CAPTCHA Utilizing Boundary Identification in AI-extended Videos

**LEHAO LIN**, The Chinese University of Hong Kong, Shenzhen, China

**KE WANG**, The Chinese University of Hong Kong, Shenzhen, China

**MAHA ABDALLAH**, Sorbonne Université, France

**WEI CAI**, University of Washington, USA

In recent years, the rapid development of artificial intelligence (AI) especially multi-modal Large Language Models (MLLMs), has enabled it to understand text, images, videos, and other multimedia data, allowing AI systems to execute various tasks based on human-provided prompts. However, AI-powered bots have increasingly been able to bypass most existing CAPTCHA systems, posing significant security threats to web applications. This makes the design of new CAPTCHA mechanisms an urgent priority. We observe that humans are highly sensitive to shifts and abrupt changes in videos, while current AI systems still struggle to comprehend and respond to such situations effectively. Based on this observation, we design and implement BounTCHA, a CAPTCHA mechanism that leverages human perception of boundaries in video transitions and disruptions. By utilizing AI's capability to expand original videos with prompts, we introduce unexpected twists and changes to create a pipeline for generating short videos for CAPTCHA purposes. We develop a prototype and conduct experiments to collect data on humans' time biases in boundary identification. This data serves as a basis for distinguishing between human users and bots. Additionally, we perform a detailed security analysis of BounTCHA, demonstrating its resilience against various types of attacks. We hope that BounTCHA will act as a robust defense, safeguarding millions of web applications in the AI-driven era.

CCS Concepts: • **Security and privacy** → **Authentication; Access control**; • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Computer vision**.

Additional Key Words and Phrases: CAPTCHA, Web Security, Automated Attacks, Human Perception, AI-extended Videos, Video Extension, Generative AI

## 1 Introduction

CAPTCHA [28, 61], an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart," is a type of test used to verify whether an online user is a human or a bot. As such, CAPTCHAs are sometimes referred to as "reverse Turing tests" [73]. Users must complete a given task and submit the result, and only after being verified as human can they proceed with further actions within an online application. Typically, CAPTCHA pop-up windows that appear when users attempt to log in or make network requests too quickly. Login page intercepts are designed to hinder bots from the start, while rapid network requests may be interpreted as potential bot activity imitating human actions.

The primary purpose of CAPTCHAs is to prevent web crawlers and bots [46] that simulate human behavior from sending frequent requests, which can make the operational data of web services more authentic, resistant to Sybil attacks [19], and place significant strain on a server's network. A large volume of such requests may be perceived as a denial-of-service (DoS) attack [39], and attackers with multiple nodes may launch a distributed denial-of-service (DDoS) attack [24, 42]. This disrupts the normal use of web applications for legitimate human users. Additionally, bots can attack social media content by promoting keywords to get them trending on X (previously Twitter), thus creating false trends that can reach a broad audience [21, 29]. The situations above are not only happened in the centralized

---

Authors' Contact Information: **Lehaol Lin**, lehaolin@link.cuhk.edu.cn, The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong, China; **Ke Wang**, kewang1@link.cuhk.edu.cn, The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong, China; **Maha Abdallah**, maha.abdallah@lip6.fr, Sorbonne Université, Paris, France; **Wei Cai**, weicaics@uw.edu, University of Washington, Tacoma, WA, USA.

web, but also and more severe in the decentralized web [48, 96]. While there are various methods to detect [41, 67, 97] and mitigate [8, 20, 36] such attacks, CAPTCHAs remain one of the most cost-effective and widely used defenses.

However, the weakness of CAPTCHA lies in its vulnerability when an attack bot gains the ability to solve its challenges, hence becoming a so-called CAPTCHA solver, and rendering it ineffective as a defense mechanism. For instance, early text-based CAPTCHAs can now be easily bypassed using simple Optical Character Recognition (OCR) tools, such as Tesseract [70]. As a result, designing new CAPTCHAs is an ongoing challenge, involving a continuous and evolving battle against increasingly sophisticated bot intelligence.

With the recent surge in multi-modal Large Language Models (MLLMs) [38, 91], artificial intelligence (AI) has gained unprecedented capabilities in handling a wide range of complex tasks, including but not limited to understanding text, multimedia, and generating synthesized conclusions. Furthermore, by granting AI the ability to control computers [54], these systems can now take autonomous actions in the digital world. An AI-powered agent bot could operate a computer much like a human, posing a significant security threat to existing web applications. Once an AI bot learns how to overcome CAPTCHA challenges, the current CAPTCHA systems could become obsolete overnight.

In recent years, AI-driven video generation models have emerged in an exponential manner, with examples such as Stable Video Diffusion (SVD) [6], Sora<sup>1</sup> [51], Pika Labs<sup>2</sup>, Runway<sup>3</sup> [17], Stability AI<sup>4</sup>, Kling<sup>5</sup>, and others [33, 34, 43, 53]. The videos generated by these models have proliferated across various social media platforms. As model architectures continue to improve and the size of model parameters expands, the length and quality of the generated videos are expected to reach new heights, making it increasingly difficult to distinguish them from real content. In addition to generating videos from prompts and related images, some models are especially capable of video extension (known as video prolongation, and video prediction). Moreover, these models can customize extended scenes, actions, visuals, and objects within the video using text prompts. They can also create visuals that do not exist in the real world, similar to special effects in films. And in the research [31], it helps explain changes in videos that violate the predictive perception are easily noticed by humans, highlighting how humans detect unnatural changes in the video content. As a result, humans are able to identify the part of a video that has been AI-extended by observing strange changes in the video.

**Motivation.** In light of the above observations, we argue that the human ability to recognize sudden events, transitions, or anomalies inconsistent with the real world in video content can be harnessed to design a novel web CAPTCHA system. This system would serve as a defense against bot attacks on the internet. The video data used for the CAPTCHA could be generated using AI techniques, specifically by leveraging the extending video capabilities of video generation models.

**Approach.** We design a novel CAPTCHA mechanism aimed at defending against current and future, more intelligent web robots and crawlers. This mechanism is based on abrupt changes and transitions in video frames and storylines, where users are asked to identify the points of transition, allowing to distinguish between human users and robots. Since collecting large datasets of real videos with significant transitions is costly, and the degree of transition in real videos is limited, we use AI video generation models to extend these transitions. As a result, our CAPTCHA mechanism relies on human perception to identify AI-extended video boundaries, and which we name BounTCHA. Based on this design, we develop a prototype whereby, after watching the video, users are asked to drag the progress bar to the point where they believe the transition occurs and submit their response. The collected timing data is used to determine the

<sup>1</sup><https://openai.com/index/sora/>

<sup>2</sup><https://pika.art/>

<sup>3</sup><https://runwayml.com/>

<sup>4</sup><https://stability.ai/>

<sup>5</sup><https://kling.kuaishou.com/>

effective range within which humans can recognize the boundary, which then serves as the basis for distinguishing human users from robots. Following this, we conduct a detailed security analysis to assess the effectiveness of this defense method.

To design a new CAPTCHA, we should follow three basic principles from [14, 26]: (1) Easy to generate and evaluate. (2) Easy for most people to solve. (3) Difficult for automated bots to solve. Therefore, we define several **Research Questions (RQs)** to investigate the BounTCHA mechanism and help to evaluate it:

**RQ1:** Is it practically feasible to create such a CAPTCHA? (Address the first principle.)

**RQ2:** To what extent can humans distinguish the boundary between the original and generated segments of the video? (Address the second principle.)

**RQ3:** Is there a potential for this CAPTCHA to be successfully attacked? What is the likelihood of such an attack occurring? (Address the third principle.)

Our **contributions** can be summarized as follows:

- To the best of our knowledge, we are the first to design and implement a novel CAPTCHA based on human perception of AI-extended video boundaries, named BounTCHA. The relevant prompts, working prototype, and videos are open-sourced on GitHub, link: <https://github.com/LehaoLin/bountcha>.
- We explore the difficulty and cost of creating this CAPTCHA, and examine the range of time biases within which humans can effectively distinguish video boundaries, demonstrating the feasibility of this system as a novel CAPTCHA.
- We conduct a security analysis of BounTCHA against various potential attack methods, including random attacks, database attacks, and multi-modal LLM attacks, and demonstrate its effectiveness in defending against robot tasks.

The following sections review relevant literature, outline video data preparation, detail the prototype and studies on human performance, discuss the in-depth security analysis and conclude with limitations and future work.

## 2 Related Work

### 2.1 CAPTCHAs

We have categorized the current mainstream types of CAPTCHA into four groups: text-based, image-based, 3D & gamified, and video CAPTCHAs.

**2.1.1 Text-based CAPTCHAs.** Text-based CAPTCHAs have various forms of representation, shown on the left part of Figure 1. These can be broadly categorized into two types: character-based and digit-based. In the character-based type, distorted text is generated for users to recognize and input the corresponding characters. The digit-based type, in addition to requiring users to recognize and input numbers, may also involve simple math arithmetic operations [32]. The correctness of the submitted answer is used to verify whether the user is human. Some websites that specifically serve certain countries or regions use their native language as elements [2, 75, 82, 92]. By overlaying text onto background images after applying filters, some provide a word or short phrase and ask the user to click on the corresponding characters in the image in the correct sequence to verify users. [89] proposes SS-CAPTCHA, which leverages the human ability to recognise strangeness in translated sentences to detect malware.

**2.1.2 Image-based CAPTCHAs.** Some of these showcases are presented in the middle and right sections of Figure 1. Common image-based CAPTCHA mechanisms include multiple image selection, jigsaw puzzles, and image position

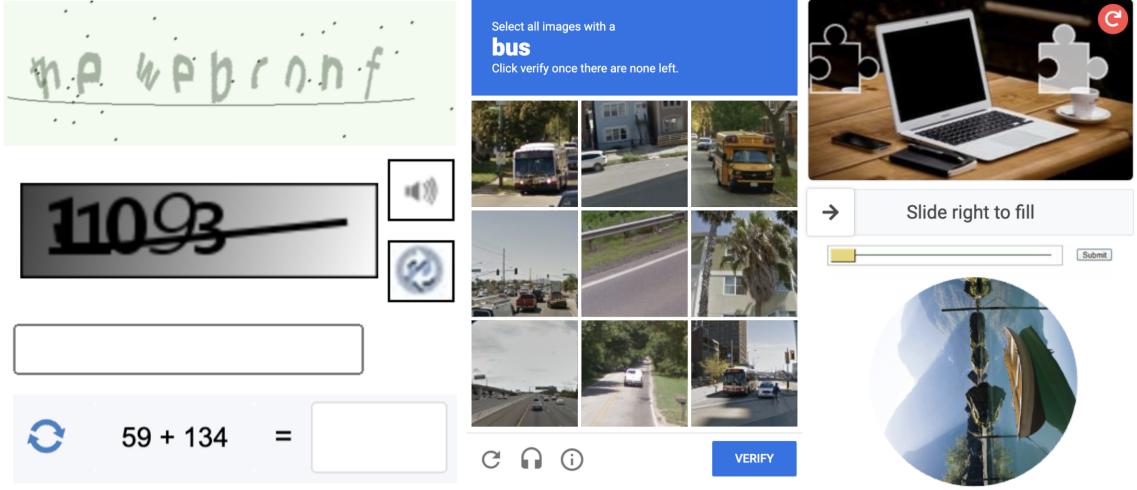


Fig. 1. Common text-based and image-based CAPTCHA examples, including arithmetic CAPTCHA, reCAPTCHA, puzzle CAPTCHA, What's Up CAPTCHA, and others.

correction. For instance, the multiple image selection used in reCAPTCHA [78] can be categorized into selecting the target object from prompts and manually performing semantic segmentation of a large image into a 3x3 grid. A more widely used jigsaw CAPTCHA [23, 62] requires the user to slide a piece into its correct position within the image. In contrast, the What's Up CAPTCHA [26] involves adjusting the placement of an image by sliding a bar. Moreover, the scene tagging [57] tests the ability to recognize a relationship among multiple objects in an image.

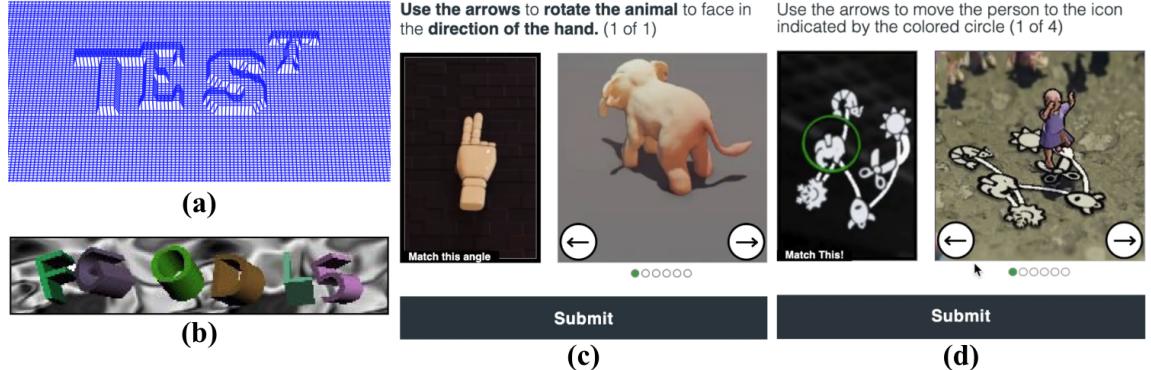


Fig. 2. Showcases of 3D & gamified CAPTCHAs. (a) is a text-based 3D CAPTCHA [60]. (b) is named 3D CAPTCHA [37]. (c) and (d) are gameified CAPTCHAs used by OpenAI's ChatGPT with the 3D view images.

**2.1.3 3D & Gamified CAPTCHAs.** Due to considerations of compatibility of 3D object rendering across different browsers and the complexity of generating 3D content, 3D CAPTCHAs often combine features of both text-based [60] and image-based [65, 83] methods. In Figure 2, (a) and (b) illustrate text-based 3D CAPTCHAs, while (c) and (d) demonstrate image-based 3D CAPTCHAs. Additionally, another mechanism, known as gamified CAPTCHAs [58],

incorporates elements of interactivity, as shown in (c) and (d), where users perform various tasks based on the instruction [1]. However, with the widespread improvement in hardware and software performance, some CAPTCHA systems opt to disregard browser compatibility and fully leverage modern features. For instance, Dotcha [44] employs dynamic scatters to represent the 3D effect.

**2.1.4 Video-based CAPTCHAs.** Similarly to our present work, there have been some studies utilizing video as an element in CAPTCHAs. For instance, [45] uses user-provided descriptions of videos to find relevant labels and tags for verification purposes. [63] improves upon this method by replacing text descriptions with a selection-based approach. [4] combines text-based CAPTCHAs with video backgrounds, requiring users to input the shown text.

**2.1.5 Drawbacks & Attacks.** Although the various CAPTCHA forms mentioned above provide varying degrees of protection for web applications, the advancement of AI capabilities has made these defenses increasingly fragile and susceptible to being bypassed [25]. For **text-based CAPTCHAs**, besides traditional Optical Character Recognition (OCR) [11] attacks, even relatively simple machine learning models or neural network architectures [76], such as Support Vector Machine (SVM), K-nearest Neighbors (KNN), and Convolutional Neural Network (CNN), can successfully carry out attacks [12, 71, 81]. For **image-based CAPTCHAs**, more complex neural network models, such as ResNet [30] or Vision Transformer (ViT) [18], can be employed for recognition. Additionally, techniques like edge detection [40, 56], object detection [64], and pixel-level segmentation [13, 52] can be applied to analyze the image, followed by a user interaction simulation programmatically to bypass these defenses [3, 22, 69, 74]. For **3D & gamified CAPTCHAs**, in addition to traditional attack methods [59], since the challenge is still displayed in the form of images on the user interface, attackers can resort to taking screenshots and treating it as an image-based CAPTCHA. They can then leverage multi-modal LLMs to better understand the task and execute the attack [7, 27, 50]. For **video-based CAPTCHAs**, there are already numerous vision language models (VLMs) [5, 55, 95, 98] capable of interpreting video content and generating textual descriptions. Moreover, this type of CAPTCHA faces challenges related to language internationalization and localization, which limits its usage to specific regions and imposes linguistic barriers, as well as difficulties in understanding complex tasks for users in different education levels.

## 2.2 Video Generative Models

Video generation models possess the ability to modify the generated content in accordance with user input, which is often known as prompts. Based on the kind of prompt, these models can be classified as text-to-video (T2V) models [15], image-to-video (I2V) models [68], and video-to-video (V2V) models [85]. Beyond the straightforward generation of videos, some I2V and V2V models enable fine-tuning, modification, editing, and even the stitching of generated videos by utilizing additional text prompts [10, 16, 35, 72].

In addition to explorations in academic research, video generation technology has been successfully applied in commercial contexts and has reached a relatively mature stage. Companies such as Kling, Runway, Pika, and Stability AI support text-to-video (T2V) and image-to-video (I2V) generation and provide corresponding API interfaces. According to a report [77] and our practical experience, Kling exhibits superior overall performance in generating videos from images, with more stable output quality. Consequently, we have decided to adopt Kling as the model for our subsequent video generation expansions.

### 2.3 Video Understanding

In the realm of video understanding tasks, video foundation models (VFs) have embarked on promising endeavors in exploring model architectures, as evidenced by references such as [47, 79, 87]. Additionally, they have made significant progress in learning paradigms, [49, 84, 86, 90, 94] serving as examples. As for the latest advancements in video understanding, Tarsier [80] emerges as a prominent family of large-scale video-language models. These models are meticulously designed to craft high-quality video descriptions. Notably, the model, along with its code and data, has been made publicly accessible for inference, evaluation, and deployment purposes. As of September 3rd, 2024, extensive evaluation results have demonstrated the remarkable capabilities of Tarsier in general video understanding. It has achieved state-of-the-art (SOTA) performance across six open benchmarks, solidifying its position as a leading approach to understand videos. Consequently, in our current research endeavor, we have decided to utilize this model to participate in the prompt generation of AI-extended videos.

### 3 Video Data Preparation (RQ1)

We collected 25 raw videos which are short (3-10 seconds) without significant changes in content from Pexels<sup>6</sup>. To avoid the unnatural plot twists commonly found in videos (such as anime) and other films as in the science fiction and fantasy genres, which could lead to user misjudgment, we do not select such videos. Instead, we choose those that are captured with real cameras, without any special effects or filters that could cause strange visual distortions.

Moreover, to ensure data diversity while avoiding overwhelming participants in the subsequent user study, which could lead to distorted data, we limit the number of selected videos. Within this constraint, we ensure that the videos feature a variety of shots, scenes, backgrounds, subjects, and other elements to make the collected data as representative as possible. In this section, we aim to answer RQ1 based on BounTCHA’s data preparation. Next, we describe details of the production pipeline for CAPTCHA video and the video properties before and after processing.

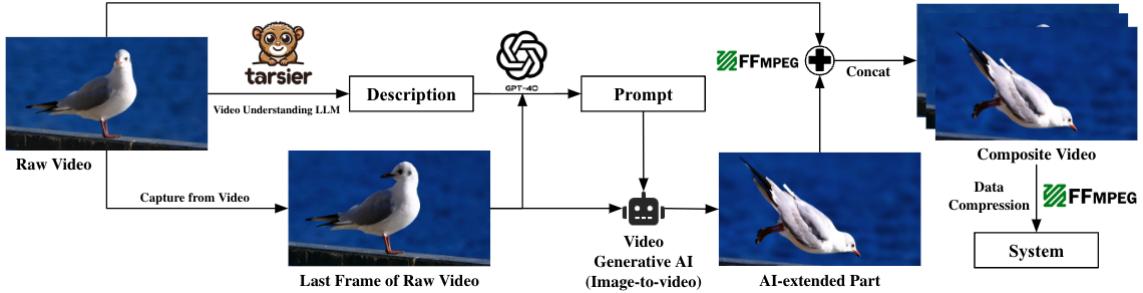


Fig. 3. The production pipeline for generating BounTCHA videos.

#### 3.1 Video Generated Pipeline

The pipeline used for CAPTCHA video production is shown in Figure 3. We first use Tarsier to understand the video content and output it as a text description. This description, along with the final frame of the video, is then input into GPT-4O to generate prompts for the video generation AI model. The prompts, along with the final frame of the raw video, are input into the video generation AI to produce the AI-extended part. Next, we merge the raw video with the

<sup>6</sup><https://pexels.com/>, where videos are free to use and allowed to be modified according to their license.

AI-extended part to create a composite video. Finally, we compress this composite video using FFmpeg to produce the video data used in the CAPTCHA system.

Since Sora is not yet available to the public by September 15th, 2024, and considering the performance and quality of the generated videos (some services do not perform well on video extension tasks), we choose Kling to handle the task of generating extended videos. In addition, given that the data size of the composite video may be too large for CAPTCHA, it could cause significant network strain on the host server in real-world scenarios. Therefore, we use FFmpeg to remove the audio track and compress the video to a target bitrate of 256k, which reduces the network transmission load while ensuring that users can still clearly view the video content.

Therefore, we can denote the raw video as  $V_{in} = \{F_1, F_2, \dots, F_n\}$ , where  $F_i$  means  $i$ th frames and  $n$  denotes the number of frames.  $LLM_v$  is the video understanding model, and  $LLM_t$  is the image-text model.  $p$  is the prompt for video understanding. Thus, the prompt for the video generative model is

$$p' = LLM_t(LLM_v(V_{in}, p), F_n) \quad (1)$$

Then, the extended part is generated as

$$V_{ext} = \{G_1, G_2, \dots, G_m\} = \text{Gen}(F_n, p') \quad (2)$$

Finally, the output after concatenation and compression is

$$V_{out} = \{F_1^*, \dots, F_n^*, G_1^*, \dots, G_m^*\} = \text{compress}(V_{in} \oplus V_{ext}) \quad (3)$$

where  $F_n^*$  can be regarded as the real boundary frame of the output video and is the target for users to identify.

### 3.2 Prompts

**3.2.1 Prompts for Content Extraction.** We provide two prompts to assist with video understanding to extract the video's content:

- **Description:** Describe the video in detail, covering all events, actions and camera motions. Also, describe the characters' appearance and the background.
- **Keywords:** Summarize the video with 5 keywords.

The video description serves to help the subsequent LLM understand the content of the video, while the five keywords enable the LLM to identify the key elements within the video. This information guides the model in determining which factors should be altered or remain unchanged for the transitions.

**3.2.2 Prompts for Video Generation Prompts.** After obtaining the description and five keywords of the raw video, we incorporate this information into the subsequent video generation prompts, along with the last frame of the raw video. These inputs are then fed into the MLLM to generate a completely new video narrative. This process ensures that the generated prompts significantly differ from the storyline and visuals of the raw video, while still being grounded in its foundational elements. Subsequently, the prompts used for video generation are input into the video generative AI to contribute to the creation of the AI-extended video.

You need to generate a prompt to instruct a video model to generate a subsequent video based on the last frame of a given natural video. You need to ensure that the expanded video differs significantly from the previous video to create a considerable difference between the raw video and the AI-extended part. The focus should be on quickly generating movements that differ from natural laws and common sense, ensuring that humans can react quickly without causing drastic changes in the visuals.

We provide:

1. **A description of the video:** {{ description from Section 3.2 }}
2. **Five keywords related to the video:** {{ keywords from Section 3.2 }}

You need to accept these Requirements:

**Requirement 1:** The prompt should satisfy: Subject + Background + Movement. **Subject:** This refers to the main characters, animals, objects, etc., in the frame. **Movement:** This indicates the desired trajectory of the target subject. **Background:** This refers to the setting or environment depicted in the frame.

**Requirement 2:** Generate only a description prompt, within 120 characters, including spaces and punctuation. No additional information is needed. Please answer in English.

**Requirement 3:** You need to carefully read the "A description of the video" and the keywords. Your expansion should be based on them. Try not to deviate too much.

### 3.3 Video Quality & Shift Cutting

To ensure that users can complete the verification process efficiently, our videos range from a minimum length of 5.12 seconds to a maximum of 14.4 seconds. Prior to compression, the smallest video was 5 MB; however, after compression, the largest video did not exceed 350 KB. To simulate a realistic CAPTCHA usage environment, we utilize the compressed videos for subsequent experiments. The sizes and lengths of the videos before and after compression are illustrated in Figure 4. Considering the practical context, to minimize the time users spend on CAPTCHA verification, we selected composite AI-extended videos with durations ranging from 5 to 15 seconds as our experimental data. Due to the inherent compression in the MPEG-4 format, the sizes of the original and compressed videos do not increase linearly with the video length. Some minor discrepancies in size and trends were observed; however, the overall trend in video sizes remains consistent. The average size of the compressed videos used is 257.68 KB, which is an acceptable size that will not place significant strain on the network bandwidth.

In addition to the operations mentioned above, we can also use video cutting to adjust the boundary position throughout the entire video duration. This allows the creation of multiple copies of the same video with different boundary positions. In real-world scenarios, this approach helps expand and enhance datasets, enabling them to handle increased traffic and calls.

### 3.4 Video Production Time Consumption

Figure 5 illustrates the time consumption in various stages of video preparation, such as video understanding for 21.7 seconds, prompt generation for 5.4 seconds, video generation for 521.2 seconds and video compression for 3.1 seconds.

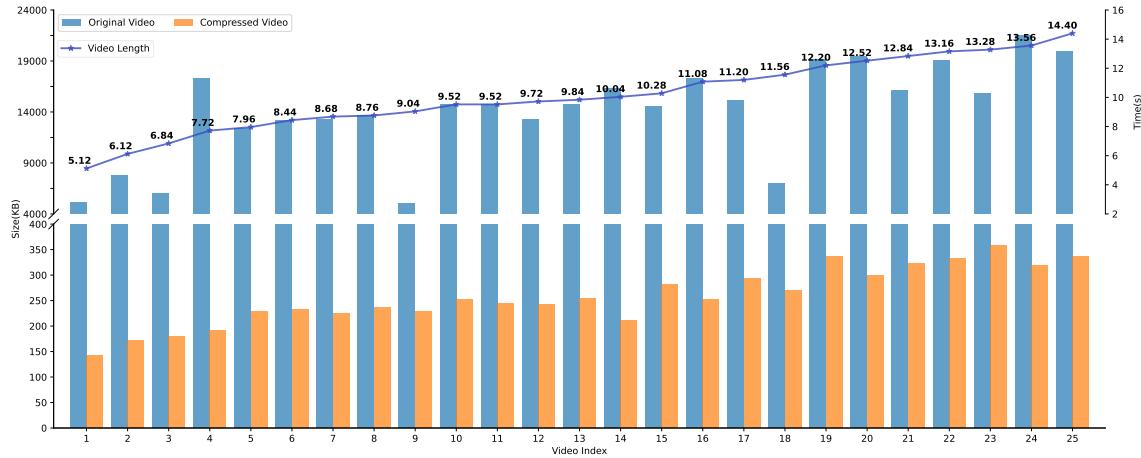


Fig. 4. A bar chart comparing the sizes of original videos ( $\mu = 14106.62$ ,  $\sigma = 4590.93$ ) and compressed videos ( $\mu = 257.68$ ,  $\sigma = 55.75$ ), alongside a line graph depicting video lengths ( $\mu = 10.00$ ,  $\sigma = 2.42$ ), with the videos indexed according to their total duration.

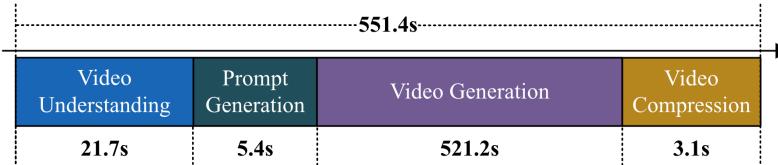


Fig. 5. The time cost of the video preparation pipeline. The length of the blocks in the figure is not drawn to scale based on the time duration.

#### 4 BounTCHA Prototype

The system prototype of BounTCHA is shown in Figure 6. The video playback area is at the top, while the user interaction area is at the bottom. Users can determine the boundary of the composite video by clicking the Play/Pause button and can also drag the slider to seek and adjust the playback. Additionally, users can see the total video length and the current playback time. When users drag the slider to the position they believe marks the boundary, they need to click the Submit button to complete the CAPTCHA. The system then transmits the video ID and the user's confirmed boundary time to the server backend, which records the user's boundary time and compares it with the actual boundary.

The architecture of the prototype is as follows: the frontend is built using Vue.js, the backend utilizes Python with FastAPI, and the database is powered by MongoDB.

In the experimental part of this work, we will gather statistics on the boundaries confirmed by users to determine the effective range of human judgment. In real CAPTCHA scenarios, the server will return the comparison result with the actual boundary. Trials that fall within the effective range are likely to have been completed by a human rather than a robot.

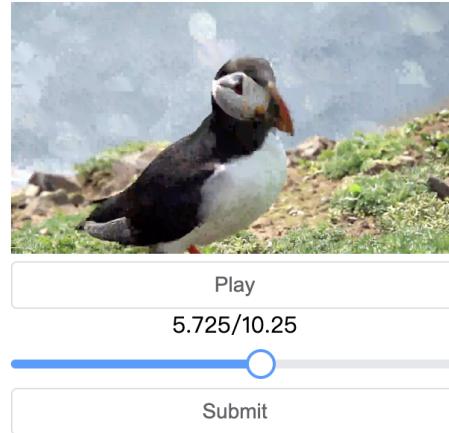


Fig. 6. User interface of the BounTCHA prototype.

## 5 Studies on Human Performance (RQ2)

### 5.1 Experiment

In this section, we conduct a user study based on the BounTCHA system to explore the discrepancy between users' perceived boundaries of AI-extended videos and the actual boundaries.

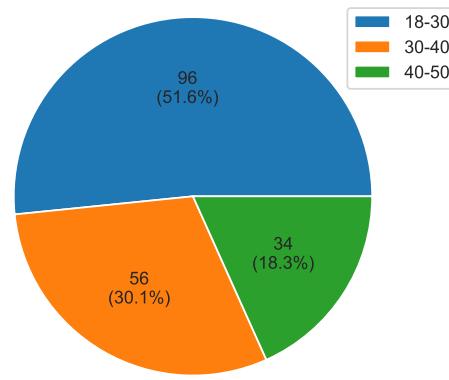


Fig. 7. The age group pie chart of the participants.

**Setup.** For ease of both offline and online experiments, we deployed the BounTCHA prototype on a remote server. The server is configured with 2 CPUs and 2GB of memory, running the Ubuntu 22.04.3 LTS operating system.

**Ethics.** The school's Institutional Review Board (IRB) reviewed and approved this human-subjects research. In order to preserve the double-blind review process, the approval number will be provided in the camera-ready version.

**Participants.** We recruited 186 participants for the experiment by posting flyers on campus and advertising on social media (comprising 83 participants in-person and 103 participants via ZOOM for online experiments). Participants'

ages ranged from 18 to 48 years ( $\mu = 26.93$ ,  $\sigma = 7.57$ ), and all had prior experience with web-based CAPTCHAs. Each participant compensated with \$0.7 USD.

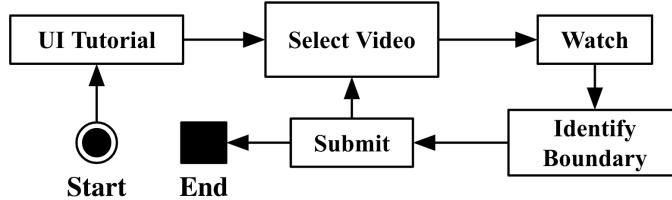


Fig. 8. The procedure of user study.

**Procedure.** As shown in Figure 8, at the beginning of the experiment, we first explained the objectives to the participants and demonstrated how to operate our experimental system. We informed the participants that the first part is the original video, while the second part is the extended video. Once the participants became familiar with the system, they sequentially completed a Boundary selection task for 25 videos (in random order). The total completion time for the 25 video experiments was approximately 10 minutes.

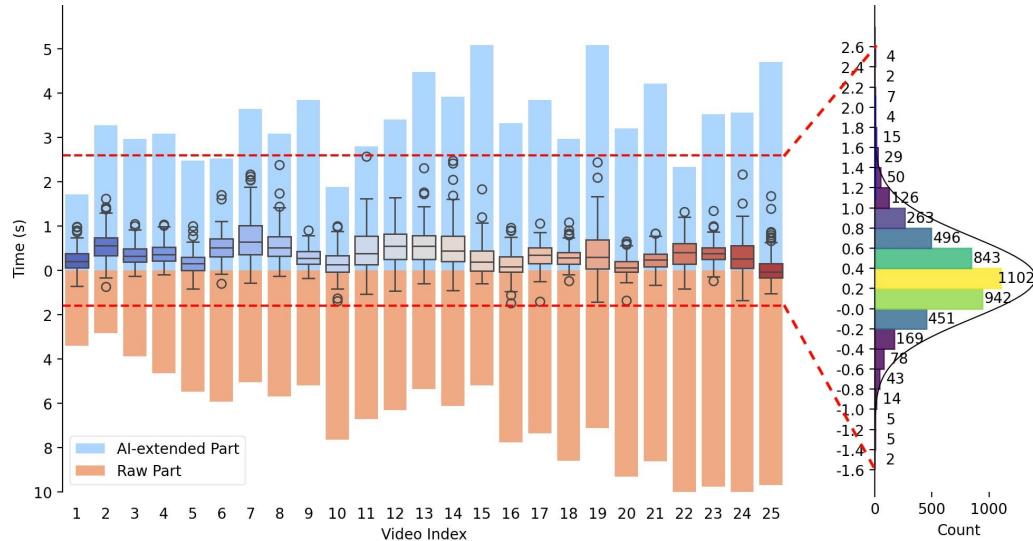


Fig. 9. Left: length of videos and distribution of time bias between the human identification boundary and the actual boundary for each video. 0 means the actual boundary between raw video and AI-extended part. Right: the count of all time bias along with its normal distribution estimate ( $\mu = 0.332$ ,  $\sigma = 0.406$ ).

## 5.2 Results

Figure 9 shows the results of the experiment. All the data of time bias falls between -1.6s and 2.6s, and it follows a normal distribution. Therefore, we can increase the difficulty of BounTCHA by adjusting the significance level to narrow the time range. In Figure 10, we show the time bias range where the confidence level varies from 0.5 to 0.95,

and the corresponding significance level varies from 0.5 to 0.05. Furthermore, to demonstrate the general applicability of the obtained time bias range, we divide the videos into 5 groups. We derive the range from 4 of the groups and use the remaining 1 group as a validation set to verify the success rate, repeating this for a total of 5 rounds with different validation sets. After the leave-one-out cross-validation, we finally calculate the average success rate, represented by the green curve in Figure 10.

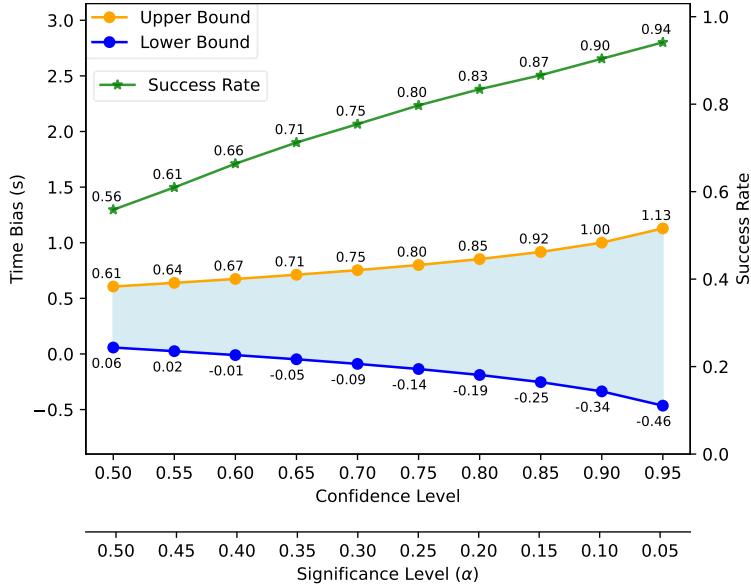


Fig. 10. The overall time bias range where the confidence levels with the corresponding significance levels. And their overall success rates from the leave-one-out cross-validation.

Based on the age groups presented in Figure 7, we also analyzed how individuals of different age ranges contributed to the overall participation and success rate. To highlight the differences in success rates among various age groups, we used the overall time bias range as the success criterion, as shown in Figure 11. From the results, we observed a subtle pattern: the time bias range tends to be slightly larger for older age groups, while their corresponding success rates are slightly lower. However, the differences between data points at the same confidence level are relatively minor. Therefore, the overall data can be used as the criterion for subsequent analysis. And this section can answer RQ2 from the user study.

Moreover, we use the Pearson correlation coefficient [66] and the Spearman's rank correlation coefficient [93] to assess how the relationship among time bias means & video lengths, and time bias standard deviations & video lengths. The calculated results all approach zero, and the p-values are very large. This indicates that there is almost no monotonic relationship between the listed variables.

## 6 Security Analysis (RQ3)

We conduct a security analysis based on three attack methods, namely the random attack, the database attack, and the multi-modal LLM attack to answer RQ3. We denote the full duration length of the video as  $L$ , the attack time bias  $x \in X$ , the mean value of the time bias as  $\mu$ , the standard deviation as  $\sigma$ , the significance level as  $\alpha$ , the two-tailed confidence

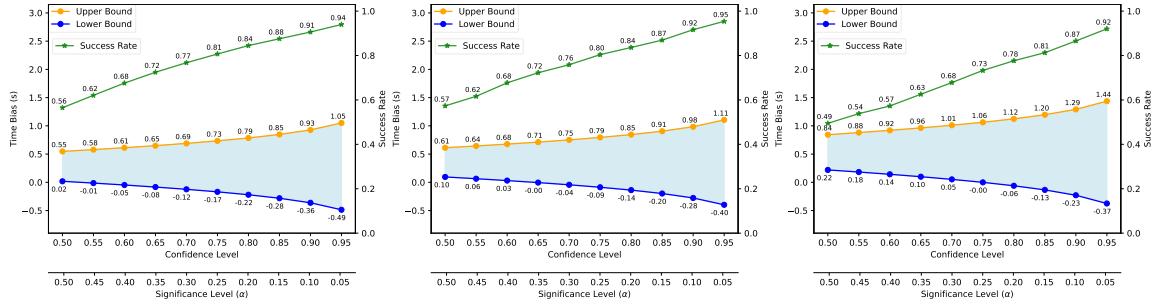


Fig. 11. Left: the time bias range and the success rate of the group in the age 18-30. Mid: in the age 30-40. Right: in the age 40-50.

Table 1. Correlation Analysis of Bias and Video Length Using Pearson and Spearman Correlation Coefficients

	Bias Mean & Video Length	Bias STD & Video Length
Pearson [66]	-0.377 ( $p = 0.062$ )	0.138 ( $p = 0.510$ )
Spearman [93]	-0.341 ( $p = 0.095$ )	0.088 ( $p = 0.675$ )

level as  $1 - \alpha$ , the percent point function as  $\text{ppf}(\cdot)$  to get z-score, and the effective time bias range  $[\beta_1, \beta_2]$ , where

$$\beta_i = \mu \pm \sigma \cdot \text{ppf}(1 - \frac{\alpha}{2}), i \in \{1, 2\} \quad (4)$$

## 6.1 Random Attack

**6.1.1 Uniform Distribution Attack.** We assume that attackers only know the video length  $L$  and use uniform random  $X \sim U(0, L)$  to attack BounTCHA. Thus, the attack success probability is

$$P(X \in [\beta_1, \beta_2]) = \int_{\beta_1}^{\beta_2} \frac{dy}{L} = \frac{\beta_2 - \beta_1}{L} = \frac{2\sigma \cdot \text{ppf}(1 - \alpha/2)}{L} \quad (5)$$

where  $\sigma$  is a constant value computed from the dataset.  $L$  and  $\alpha$  are variables. Therefore, we can draw a chart to show the relationship among  $\alpha$ ,  $L$ , and  $P(X \in [\beta_1, \beta_2])$ , shown as Figure 12.

From the figure, we observe that as  $\alpha$  and  $L$  increase, the success rate of the attack decreases. Notably, within the range where  $\alpha$  is between 0.05 and 0.25 and  $L$  is between 5 and 7.5, the success rate declines rapidly. Beyond this range, the rate of decline tends to stabilize. Therefore, to defend against uniform distribution random attacks, it is recommended to set  $\alpha \geq 0.25$  and  $L \geq 7.5$ .

**6.1.2 Truncated Normal Distribution Attack.** We assume that attackers only know the video length  $L$  and use the two-tailed truncated normal distribution random to attack BounTCHA, where  $\mu' = \frac{L}{2}$ , and its probability density function (PDF) is

$$f(x; \mu', \sigma', 0, L) = \frac{1}{\sigma'} \frac{\varphi(\frac{x-L/2}{\sigma'})}{\Phi(\frac{L/2}{\sigma'}) - \Phi(\frac{-L/2}{\sigma'})}, 0 \leq x \leq L \quad (6)$$

$$\varphi(\xi) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\xi^2), \Phi(x) = \frac{1}{2}(1 + \text{erf}(x/\sqrt{2})) \quad (7)$$

and  $f = 0$  otherwise, where  $\text{erf}(\cdot)$  is the Gauss error function

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2) dt \quad (8)$$

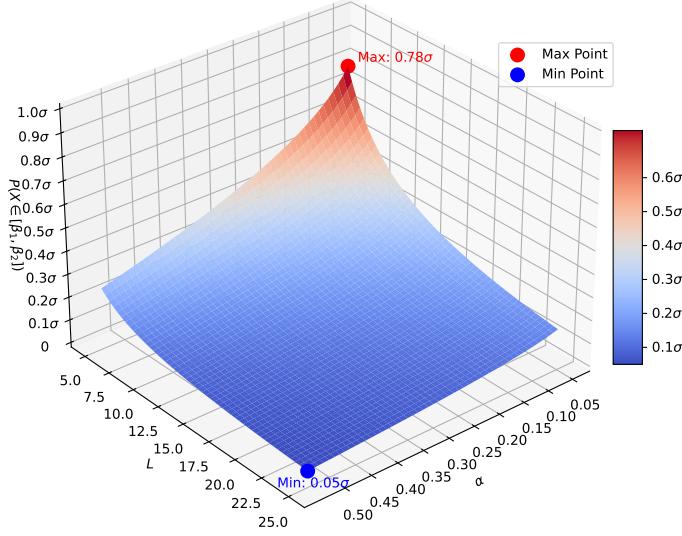


Fig. 12. 3D Surface Plot of  $P(X \in [\beta_1, \beta_2])$ : Visualization of the  $P(X \in [\beta_1, \beta_2])$  in respect to variables  $L$  and  $\sigma$ .

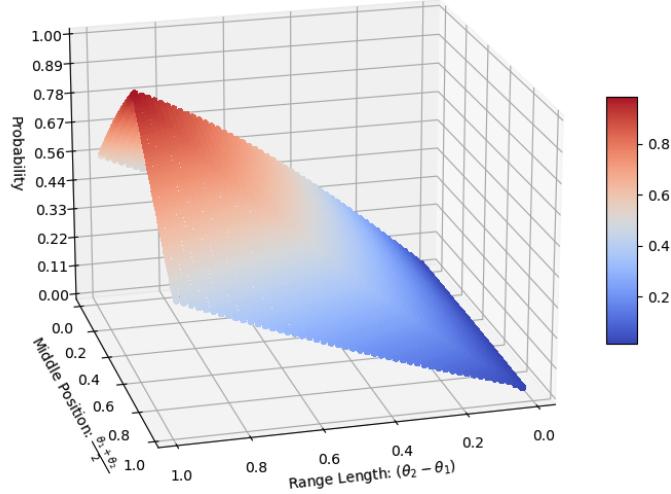


Fig. 13. 3D Surface Plot of truncated normal distribution random attack success probability in respect to the time bias range proportion and position.

Let  $\theta_1 = \frac{\beta_1}{L}$  and  $\theta_2 = \frac{\beta_2}{L}$  denote proportions of the upper bound and the lower bound with respect to the whole video length, where  $0 \leq \theta_1 \leq \theta_2 \leq 1$ . So  $\theta_2 - \theta_1$  can represent the time bias range length, and  $\frac{\theta_1 + \theta_2}{2}$  can represent the middle

position in the whole length. The attack success probability is

$$\begin{aligned} P(X \in [\beta_1, \beta_2]) &= F(\beta_2; \mu', \sigma', 0, L) - F(\beta_1; \mu', \sigma', 0, L) \\ &= F(\theta_2; \frac{1}{2}, \sigma'', 0, 1) - F(\theta_1; \frac{1}{2}, \sigma'', 0, 1) \end{aligned} \quad (9)$$

where  $F(\cdot)$  is the cumulative distribution function (CDF).

As shown in Figure 13, it is evident that the closer the midpoint of the time bias is to the midpoint of the video duration, the higher the success rate of the truncated normal distribution random attack. Additionally, the shorter the range length proportion, the lower the success rate of bot attacks. From the chart, we can deduce that  $\frac{\theta_1+\theta_2}{2} \leq 0.4$  or  $\frac{\theta_1+\theta_2}{2} \geq 0.6$  with  $\theta_2 - \theta_1 \leq 0.5$  is a suitable configuration for the video production.

## 6.2 Database Attack

Inspired by the analysis in the work [65], we discuss scenarios where an attacker has partial knowledge of the video database. For example, the attacker might have limited access to certain videos or some understanding of the general boundary distribution in the video. This consideration allows us to assess the robustness of the proposed BounTCHA system and identify potential vulnerabilities under such informed attack scenarios.

Since we utilize a video cutting method that alters video boundary positions, a single video can have multiple variants. The set of all variants from one single video is referred to as "group". The entire video database is composed of several groups of such variants. Therefore, a BounTCHA video may fall into one of the following three scenarios: (1) the attacker knows the video and its boundary; (2) the attacker does not know this particular video, but is familiar with other variants in the same group and their boundaries; or (3) the attacker is unfamiliar with both the video and any other variants within the same group, as well as their boundaries.

In these three scenarios, the attacker's success probability differs. By comparing frames of the video, the attacker can determine whether the video is fully known. If the video is known, the attacker can leverage existing boundary information to successfully launch an attack. For unknown videos, the attacker can only rely on guesswork. However, when the attacker has knowledge of other variants within the same group, she/he can rule out certain possibilities, thereby increasing the likelihood of a successful guess.

We denote the total number of groups as  $M$ , with  $m$  groups containing at least one variant known by the attacker. Every video group  $i$  has  $U_i$  variants, where  $i \in \{1, \dots, M\}$ . In the  $i$ -th group, the number of known videos is  $u_i$ , and  $\gamma_i$  refers to the attack success probability. Additionally, we use  $\gamma_0$  to represent the success probability that is fully based on guesswork.

The probability  $\sum_{i=1}^M \gamma_i$  of the bot successful attack is

$$P = \frac{\sum_{i=1}^m u_i}{\sum_{i=1}^M U_i} + \frac{\sum_{i=1}^m [(U_i - u_i)\gamma_i]}{\sum_{i=1}^M U_i} + \frac{\sum_{i=m+1}^M U_i}{\sum_{i=1}^M U_i} \gamma_0 \quad (10)$$

Furthermore, we find that  $\gamma_i$  is correlated to  $u_i$  and  $U_i$ , which means that if the attacker knows about other variants in the group, the attack success rate is higher. Additionally, we assume that variants' boundaries within one group are uniformly distributed over the video length. Thus,  $\gamma_i = \frac{1}{U_i - u_i}$ , and the attack success probability is

$$P = \frac{mu}{MU} + \frac{1}{U} \quad (11)$$

Therefore, we formulate the attack probability with  $k$  success times in the coming  $n$  rounds with the following binomial distribution:

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (12)$$

We assume that the attacker has launched 1,000 attacks. According to the different proportions of data in possession of the attacker to the total data, we have plotted the probability density distribution of the number of successful attacks in Figure 14, considering the proportion of the groups containing at least one known variant ( $\omega_1$ ), and the proportion of the total number of video variants known by the attacker across all groups ( $\omega_2$ ). Thus,  $P = \omega_1 \cdot \omega_2 + \frac{1}{U}$ , where  $\omega_1$  and  $\omega_2$  are symmetrical. From the examination of each subfigure, our findings indicate that as the values of either  $\omega_1$  or  $\omega_2$  increase, the number of successful attacks also rises. Additionally, as the curve approaches the extreme values of the x-axis range, the probability distribution becomes narrower and steeper, which means attacks are more stable and have less fluctuation in the situation. Comparing the differences between the left and right subfigures, we observe that as  $U$  increases, the probability distribution shifts leftward. This suggests that as the number of variants within the group grows, the number of successes decreases, which aligns with both common sense and our intuition.

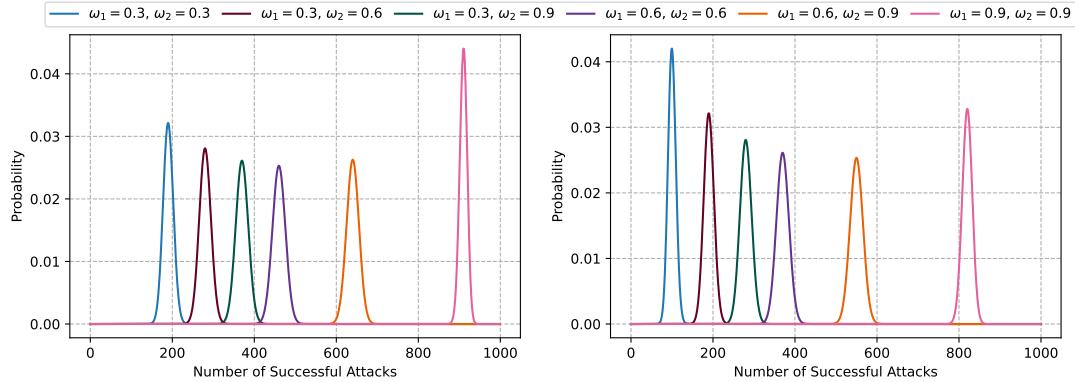


Fig. 14. Left: Probability density with  $M = 1000$  and  $U = 10$ , where  $\omega_1 = \frac{m}{M}$  and  $\omega_2 = \frac{u}{U}$  ranging from 0.3 to 0.9. Alongside the number of attacks, it shows the corresponding probabilities. Right: Similar to the left, with  $M = 1000$  and  $U = 100$ .

### 6.3 Multi-modal LLM (MLLM) Attack

We tested the recognition capabilities of two open-source MLLMs, Tarsier [80] and MiniVPM-V 2.6 [90], for the boundary of AI-extended Videos. Tarsier is at the state-of-the-art (SOTA) in multiple video question answering benchmarks. MiniVPM-V 2.6 outperforms commercial closed-source models such as GPT-4V, Claude 3.5 Sonnet, and LLaVA-NeXT-Video-34B in Video-MME performance.

We conducted three rounds of tests on the two models (Tarsier and MiniVPM-V 2.6) as well as two commercial closed-source models (GPT-4V and Claude 3.5 Sonnet) using our 25 videos. As shown in Table 2, in a total of 75 tests, the

<sup>7</sup>The relevant data is from Figure 10.

Table 2. The success rate and inference time of the two models for the recognition task.

	Success Rate	Time (Mean)	Time (Worst)
Tarsier-34b	13.33% (10/75)	20.9s	38.3s
MiniVPM-V 2.6	17.33% (13/75)	24.7s	37.9s
GPT-4V	9.33% (7/75)	26.4s	43.2s
Claude 3.5 Sonnet	10.67% (8/75)	21.2s	37.1s
Human	$\geq 82\% (\alpha \leq 0.25)^7$	14.2s	26.6s

accuracy rates of both are no more than 20%, and the running times are all over 20 seconds. The test found that current multimodal large models cannot accurately complete the task of identifying the boundary of AI-extended videos. This is manifested in: inability to understand the task given by the prompt words and returning the frame where the boundary is located instead of video seconds; in some cases, the given video seconds are too different from the real boundary, and in some cases, even the given seconds exceed the total length of the video; there is randomness, as the same prompt words and video will return very different results in multiple rounds of tests.

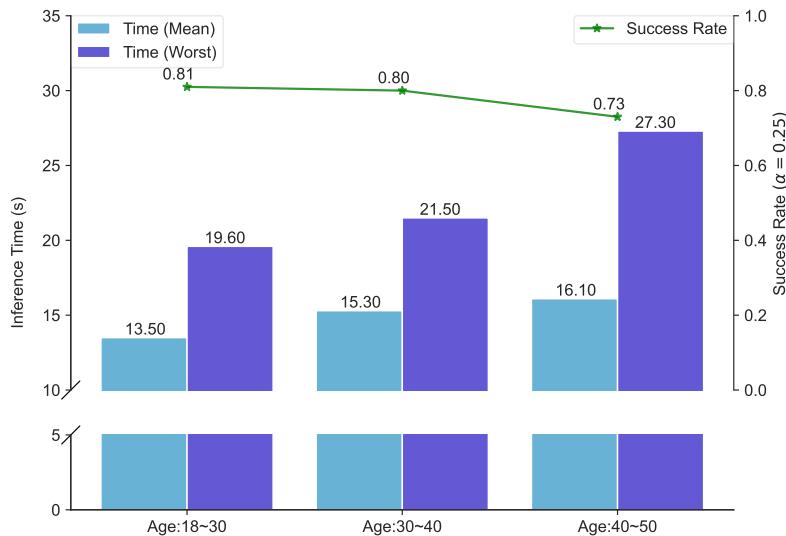


Fig. 15. The chart illustrating the relationship between age groups and the average time spent, the longest time spent, and success rates when  $\alpha = 0.25$ .

The time required to complete a verification code is an aspect of human-machine verification. On average, it takes around 10 seconds for a person to solve a typical CAPTCHA [9, 88]. As shown in Table 2, the average inference time of MLLMs all exceeds 20 seconds. In our experiment, the average time for humans to complete a task is 14.2 seconds, showing a remarkable difference among them. To provide a better indicator of real-world usability, in addition to the average time, we also recorded worst-case measurements, such as the longest verification time. The human group still exhibited the shortest verification times among all.

Moreover, MLLMs also have high costs. Running MLLMs requires a graphics card with more than 16GB of video memory, which has very high requirements for the attacker's equipment.

The chart in Figure 15 illustrates the relationship between the average time spent, the time spent on the worst-case scenario, and the success rate across age groups of human participants. We observed that older age groups required more time (both average and worst-case times) and had a lower success rate. The oldest group, in particular, had the longest time spent, which was also notably higher than the worst-case time for other MLLMs. This suggests that human users can currently be distinguished from MLLMs based on time spent when solving BounTCHA. Additionally, the success rates of human users were significantly higher than those of MLLM solvers.

## 7 Limitations and Future Work

In this section, we point out some limitations of this work, and identify future research directions.

**Participants' diversity.** Since the participants we recruited are from a university setting, they are either currently pursuing or have already completed higher education, which may result in a lack of diversity among the participants. Future research could expand the range of participants' educational levels to better assess the generalizability of this type of CAPTCHA.

**Bidirectional extension.** The mainstream direction of video extension is forward-only, but Sora supports bidirectional extensions<sup>8</sup>, which include both forward and backward. Nevertheless, Sora is currently not available to the public. Research can be conducted on bidirectional extended videos without telling users which part is generated. We can also design the derived CAPTCHA mechanisms from it.

## 8 Conclusion

We designed and implemented a novel CAPTCHA, named BounTCHA, to address the growing threat posed by increasingly intelligent AI-powered bots. BounTCHA was designed to differentiate between genuine users and bots based on human perception and identification of time boundaries in video transitions and abrupt changes. We developed a prototype of BounTCHA to conduct experiments and determine the effective range of human perceptual time bias, which serves as a basis for distinguishing between real users and bots. A comprehensive security analysis was then performed on BounTCHA, covering random attacks, database attacks, and multi-modal LLM attacks. The results of the analysis demonstrated that BounTCHA effectively defends against various attack vectors. We envision BounTCHA as a robust shield in web security, safeguarding millions of web applications from AI-powered bot threats in an era where machines are becoming increasingly intelligent.

## References

- [1] Suhas Aggarwal. 2013. Animated CAPTCHAs and games for advertising. In *Proceedings of the 22nd International Conference on World Wide Web*. 1167–1174.
- [2] YeonChan Ahn, Namsoo Kim, and Yoo-Sung Kim. 2013. A user-friendly image-text fusion CAPTCHA for secure web services. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*. 550–554.
- [3] Fatmawati Alqahtani and Fawaz A Alsulaiman. 2020. Is image-based CAPTCHA secure against attacks based on machine learning? An experimental study. *Computers & Security* 88 (2020), 101635.
- [4] K Anjitha and IK Rijin. 2015. Captcha as graphical passwords-enhanced with video-based captcha for secure services. In *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 213–217.
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. (2023).
- [6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

---

<sup>8</sup><https://openai.com/index/video-generation-models-as-world-simulators/>

- [7] Neha Pradyumna Bora and Dinesh Chandra Jain. 2023. A web authentication biometric 3D animated CAPTCHA system using artificial intelligence and machine learning approach. *Journal of Artificial Intelligence and Technology* 3, 3 (2023), 126–133.
- [8] Jose Brustoloni. 2002. Protecting electronic commerce from distributed denial-of-service attacks. In *Proceedings of the 11th international conference on World Wide Web*. 553–561.
- [9] Elie Bursztein, Steven Bethard, Celine Fabry, John C Mitchell, and Dan Jurafsky. 2010. How good are humans at solving CAPTCHAs? A large scale evaluation. In *2010 IEEE symposium on security and privacy*. IEEE, 399–413.
- [10] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23040–23050.
- [11] Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, Soumya K Ghosh, Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, and Soumya K Ghosh. 2017. *Optical character recognition systems*. Springer.
- [12] Jun Chen, Xiangyang Luo, Yanqing Guo, Yi Zhang, and Daofu Gong. 2017. A Survey on Breaking Technique of Text-Based CAPTCHA. *Security and communication networks* 2017, 1 (2017), 6898617.
- [13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.
- [14] Monica Chew and J Doug Tygar. 2004. Image recognition captchas. In *International Conference on Information Security*. Springer, 268–279.
- [15] Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. 2024. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131* (2024).
- [16] Paul Couairon, Clément Rambour, Jean-Emmanuel Haugéard, and Nicolas Thome. 2023. Videdit: Zero-shot and spatially aware text-driven video editing. *Transactions on Machine Learning Research* (2023).
- [17] Yifan Cui, Xinyi Shan, and Jeanhun Chung. 2024. A Feasibility Study on RUNWAY GEN-2 for Generating Realistic Style Images. *International Journal of Internet, Broadcasting and Communication* 16, 1 (2024), 99–105.
- [18] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [19] John R Douceur. 2002. The sybil attack. In *International workshop on peer-to-peer systems*. Springer, 251–260.
- [20] Zhengjie Du, Yuekang Li, Yaowen Zheng, Xiaohan Zhang, Cen Zhang, Yi Liu, Sheikh Mahbub Habib, Xinghua Li, Linzhong Wang, Yang Liu, et al. 2024. Medusa: Unveil Memory Exhaustion DoS Vulnerabilities in Protocol Implementations. In *Proceedings of the ACM on Web Conference 2024*. 1668–1679.
- [21] Tuğrulcan Elmas. 2023. Analyzing activity and suspension patterns of twitter bots attacking turkish twitter trends by a longitudinal dataset. In *Companion Proceedings of the ACM Web Conference 2023*. 1404–1412.
- [22] Christoph Fritsch, Michael Netter, Andreas Reisser, and Günther Pernul. 2010. Attacking Image Recognition Captcha s: A Naive but Effective Approach. In *Trust, Privacy and Security in Digital Business: 7th International Conference, TrustBus 2010, Bilbao, Spain, August 30-31, 2010. Proceedings* 7. Springer, 13–25.
- [23] Haichang Gao, Dan Yao, Honggang Liu, Xiyang Liu, and Liming Wang. 2010. A novel image based CAPTCHA using jigsaw puzzle. In *2010 13th IEEE international conference on computational science and engineering*. IEEE, 351–356.
- [24] Paolo Gasti, Gene Tsudik, Ersin Uzun, and Lixia Zhang. 2013. DoS and DDoS in named data networking. In *2013 22nd International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–7.
- [25] Nethanel Gelernter and Amir Herzberg. 2016. Tell me about yourself: The malicious captcha attack. In *Proceedings of the 25th International Conference on World Wide Web*. 999–1008.
- [26] Rich Gossweiler, Maryam Kamvar, and Shumeet Baluja. 2009. What's up CAPTCHA? A CAPTCHA based on image orientation. In *Proceedings of the 18th international conference on World wide web*. 841–850.
- [27] Gilad Gressel, Rahul Pankajakshan, and Yisroel Mirsky. 2024. Discussion Paper: Exploiting LLMs for Scam Automation: A Looming Threat. In *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*. 20–24.
- [28] Meriem Guerar, Luca Verderame, Mauro Migliardi, Francesco Palmieri, and Alessio Merlo. 2021. Gotta CAPTCHA'Em all: a survey of 20 Years of the human-or-computer Dilemma. *ACM Computing Surveys (CSUR)* 54, 9 (2021), 1–33.
- [29] Chris Hays, Zachary Schutzman, Manish Raghavan, Erin Walk, and Philipp Zimmer. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for Twitter bot detection. In *Proceedings of the ACM web conference 2023*. 3660–3669.
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [31] Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. 2019. Perceptual straightening of natural videos. *Nature neuroscience* 22, 6 (2019), 984–991.
- [32] Carlos Javier Hernandez-Castro and Arturo Ribagorda. 2010. Pitfalls in CAPTCHA design and implementation: The Math CAPTCHA, a case study. *computers & security* 29, 1 (2010), 141–157.
- [33] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).
- [34] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.
- [35] Yaosi Hu, Chong Luo, and Zhenzhong Chen. 2022. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18219–18228.

- [36] Guodong Huang, Chuan Ma, Ming Ding, Yuwen Qian, Chunpeng Ge, Liming Fang, and Zhe Liu. 2023. Efficient and low overhead website fingerprinting attacks and defenses based on TCP/IP traffic. In *Proceedings of the ACM Web Conference 2023*. 1991–1999.
- [37] Montree Imsamai and Suphakant Phimoltares. 2010. 3D CAPTCHA: A next generation of the CAPTCHA. In *2010 International Conference on Information Science and Applications*. IEEE, 1–8.
- [38] Iat Long Iong, Xiao Liu, Yuxuan Chen, Hanyu Lai, Shuntian Yao, Pengbo Shen, Hao Yu, Yuxiao Dong, and Jie Tang. 2024. OpenWebAgent: An Open Toolkit to Enable Web Agents on Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 72–81.
- [39] Rutvij H Jhaveri, Sankita J Patel, and Devesh C Jinwala. 2012. DoS attacks in mobile ad hoc networks: A survey. In *2012 second international conference on advanced computing & communication technologies*. IEEE, 535–541.
- [40] Junfeng Jing, Shenjuan Liu, Gang Wang, Weichuan Zhang, and Changming Sun. 2022. Recent advances on image edge detection: A comprehensive review. *Neurocomputing* 503 (2022), 259–271.
- [41] Hongwen Kang, Kuansan Wang, David Soukal, Fritz Behr, and Zijian Zheng. 2010. Large-scale bot detection for search engines. In *Proceedings of the 19th international conference on World wide web*. 501–510.
- [42] Mohammad Karami, Youngsam Park, and Damon McCoy. 2016. Stress testing the booters: Understanding and undermining the business of DDoS services. In *Proceedings of the 25th International Conference on World Wide Web*. 1033–1043.
- [43] Levon Khachatryan, Andranik Mousisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15954–15964.
- [44] Suzi Kim and Sunghee Choi. 2019. Dotcha: A 3d text-based scatter-type captcha. In *Web Engineering: 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11–14, 2019, Proceedings 19*. Springer, 238–252.
- [45] Kurt Alfred Kluever and Richard Zanibbi. 2009. Balancing usability and security in a video CAPTCHA. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. 1–11.
- [46] Karel Kubicek, Jakob Merane, Ahmed Bouhoula, and David Basin. 2024. Automating Website Registration for Studying GDPR Compliance. In *Proceedings of the ACM on Web Conference 2024*. 1295–1306.
- [47] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7331–7341.
- [48] Jiangtao Li, Ninghui Li, XiaoFeng Wang, and Ting Yu. 2009. Denial of service attacks and defenses in decentralized trust management. *International Journal of Information Security* 8 (2009), 89–101.
- [49] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. 2023. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23119–23129.
- [50] Qiujie Li. 2015. A computer vision attack on the ARTiFACIAL CAPTCHA. *Multimedia Tools and Applications* 74 (2015), 4583–4597.
- [51] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177* (2024).
- [52] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.
- [53] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320* (2023).
- [54] Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2024. CoCo-Agent: A Comprehensive Cognitive MLLM Agent for Smartphone GUI Automation. In *Findings of the Association for Computational Linguistics ACL 2024*. 9097–9110.
- [55] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).
- [56] Raman Maini and Himanshu Aggarwal. 2009. Study and comparison of various image edge detection techniques. *International journal of image processing (IJIP)* 3, 1 (2009), 1–11.
- [57] Peter Matthews, Andrew Mantel, and Cliff C Zou. 2010. Scene tagging: image-based CAPTCHA using image composition and object relationships. In *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*. 345–350.
- [58] Manaf Mohamed, Niharika Sachdeva, Michael Georgescu, Song Gao, Nitesh Saxena, Chengui Zhang, Ponnurangan Kumaraguru, Paul C Van Oorschot, and Wei-Bang Chen. 2014. A three-way investigation of a game-captcha: automated attacks, relay attacks and usability. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*. 195–206.
- [59] Vu Duc Nguyen, Yang-Wai Chow, and Willy Susilo. 2012. Breaking a 3D-based CAPTCHA scheme. In *Information Security and Cryptology-ICISC 2011: 14th International Conference, Seoul, Korea, November 30-December 2, 2011. Revised Selected Papers 14*. Springer, 391–405.
- [60] Vu Duc Nguyen, Yang-Wai Chow, and Willy Susilo. 2014. On the security of text-based 3D CAPTCHAs. *Computers & security* 45 (2014), 84–99.
- [61] Behzad Ousat, Esteban Schafir, Duc C Hoang, Mohammad Ali Tofighi, Cuong V Nguyen, Sajjad Arshad, Selcuk Uluagac, and Amin Kharraz. 2024. The Matter of Captchas: An Analysis of a Brittle Security Feature on the Modern Web. In *Proceedings of the ACM on Web Conference 2024*. 1835–1846.
- [62] Nitisha Payal, Nidhi Chaudhary, and Parma Nand Astya. 2012. JigCAPTCHA: An Advanced Image-Based CAPTCHA Integrated with Jigsaw Piece Puzzle using AJAX. *International Journal of Soft Computing and Engineering (IJSCE)* 2, 5 (2012), 2231–2307.

- [63] M Kameswara Rao, MSVK Maniraj, and B Sneha Ganga. 2014. Improved video captcha. *Journal of Emerging Technologies in Web Intelligence* 6, 4 (2014), 416–416.
- [64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.
- [65] Steven A Ross, J Alex Halderman, and Adam Finkelstein. 2010. Sketcha: a captcha based on line drawings of 3d models. In *Proceedings of the 19th international conference on World wide web*. 821–830.
- [66] Philip Sedgwick. 2012. Pearson’s correlation coefficient. *Bmj* 345 (2012).
- [67] Asuman Senol, Alisha Ukani, Dylan Cutler, and Igor Bilogrevic. 2024. The Double Edged Sword: Identifying Authentication Pages and their Fingerprinting Behavior. In *Proceedings of the ACM on Web Conference 2024*. 1690–1701.
- [68] Dongyu She and Kun Xu. 2022. An image-to-video model for real-time video enhancement. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1837–1846.
- [69] Suphanee Sivakorn, Iasonas Polakis, and Angelos D Keromytis. 2016. I am robot(deep) learning to break semantic image captchas. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 388–403.
- [70] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633.
- [71] Oleg Starostenko, Claudia Cruz-Perez, Fernando Uceda-Ponga, and Vicente Alarcon-Aquino. 2015. Breaking text-based CAPTCHAs with variable word and character orientation. *Pattern Recognition* 48, 4 (2015), 1101–1112.
- [72] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. 2024. Diffusion Model-Based Video Editing: A Survey. *arXiv preprint arXiv:2407.07111* (2024).
- [73] Mayumi Takaya, Yusuke Tsuruta, and Akihiro Yamamura. 2013. Reverse Turing Test using Touchscreens and CAPTCHA. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.* 4, 3 (2013), 41–57.
- [74] Mengyun Tang, Haichang Gao, Yang Zhang, Yi Liu, Ping Zhang, and Ping Wang. 2018. Research on deep learning techniques in breaking text-based captchas and designing image-based captcha. *IEEE Transactions on Information Forensics and Security* 13, 10 (2018), 2522–2537.
- [75] Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. 2021. Development and Evaluation of Swahili Text Based CAPTCHA. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 293–297.
- [76] Sheng Tian and Tao Xiong. 2020. A generic solver combining unsupervised learning and representation learning for breaking text-based captchas. In *Proceedings of The Web Conference 2020*. 860–871.
- [77] Upthrust. 2024. Runway, Luma, Kling, Pika, and Haiper: AI Video Generators Review Roundup. (August 2024). <https://upthrust.co/2024/08/runway-luma-kling-pika-and-haiper-ai-video-generators-review-roundup> Accessed: 2024-10-02.
- [78] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. recaptcha: Human-based character recognition via web security measures. *Science* 321, 5895 (2008), 1465–1468.
- [79] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6598–6608.
- [80] Jiawei Wang, Liping Yuan, and Yuchen Zhang. 2024. Tarsier: Recipes for Training and Evaluating Large Video Description Models. *arXiv preprint arXiv:2407.00634* (2024).
- [81] Ping Wang, Haichang Gao, Xiaoyan Guo, Chenxuan Xiao, Fuqi Qi, and Zheng Yan. 2023. An experimental investigation of text-based captcha attacks and their robustness. *Comput. Surveys* 55, 9 (2023), 1–38.
- [82] Tingting Wang and Jørgen Bøegh. 2014. Multi-layer CAPTCHA based on Chinese character deformation. In *Trustworthy Computing and Services: International Conference, ISCTCS 2013, Beijing, China, November 2013, Revised Selected Papers*. Springer, 205–211.
- [83] Simon S Woo, Jingul Kim, Duoduo Yu, and Beomjun Kim. 2017. Exploration of 3D texture and projection for new CAPTCHA design. In *Information Security Applications: 17th International Workshop, WISA 2016, Jeju Island, Korea, August 25–27, 2016, Revised Selected Papers 17*. Springer, 353–365.
- [84] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13204–13214.
- [85] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2023. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647* (2023).
- [86] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084* (2021).
- [87] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. *arXiv:2404.16994 [cs.CV]* <https://arxiv.org/abs/2404.16994>
- [88] Xin Xu, Lei Liu, and Bo Li. 2020. A survey of CAPTCHA technologies to distinguish between human and computer. *Neurocomputing* 408 (2020), 292–307.
- [89] Takumi Yamamoto, J Doug Tygar, and Masakatsu Nishigaki. 2010. CAPTCHA using strangeness in machine translation. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications*. IEEE, 430–437.
- [90] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).

- [91] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).
- [92] Junnan Yu, Xuna Ma, and Ting Han. 2017. Usability investigation on the localization of text captchas: take chinese characters as a case study. In *Transdisciplinary Engineering: A Paradigm Shift*. IOS Press, 233–242.
- [93] Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics* 7 (2005).
- [94] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems* 34 (2021), 23634–23651.
- [95] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [96] Shijie Zhang and Jong-Hyouk Lee. 2019. Double-spending with a sybil attack in the bitcoin decentralized network. *IEEE transactions on Industrial Informatics* 15, 10 (2019), 5715–5722.
- [97] Ziyi Zhang, Shuofei Zhu, Jaron Mink, Aiping Xiong, Linhai Song, and Gang Wang. 2022. Beyond bot detection: combating fraudulent online survey takers. In *Proceedings of the ACM Web Conference 2022*. 699–709.
- [98] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.