MCA-Bench: A Multimodal Benchmark for Evaluating CAPTCHA Robustness Against VLM-based Attacks

Zonglin Wu[1], Yule Xue[1], Yaoyao Feng[1], Xiaolong Wang[1], Yiren Song[2]*

[1] Southwest University
[2] National University of Singapore

Abstract

As automated attack techniques rapidly advance, CAPTCHAs remain a critical defense mechanism against malicious bots. However, existing CAPTCHA schemes encompass a diverse range of modalities—from static distorted text and obfuscated images to interactive clicks, sliding puzzles, and logic-based questions—yet the community still lacks a unified, large-scale, multimodal benchmark to rigorously evaluate their security robustness. To address this gap, we introduce MCA-Bench, a comprehensive and reproducible benchmarking suite that integrates heterogeneous CAPTCHA types into a single evaluation protocol. Leveraging a shared vision–language model backbone, we fine-tune specialized cracking agents for each CAPTCHA category, enabling consistent, cross-modal assessments. Extensive experiments reveal that MCA-Bench effectively maps the vulnerability spectrum of modern CAPTCHA designs under varied attack settings, and—crucially—offers the first quantitative analysis of how challenge complexity, interaction depth, and model solvability interrelate. Based on these findings, we propose three actionable design principles and identify key open challenges, laying the groundwork for systematic CAPTCHA hardening, fair benchmarking, and broader community collaboration.

Code — https://github.com/noheadwuzonglin/MCA-Bench

Datasets — https://www.kaggle.com/datasets/luffy798/mca-benchmultimodal-captchas

Introduction

With rapid advances in AI security—such as adversarial attacks (Song et al. 2025, 2024; Zhang et al. 2025; Ke et al. 2025), digital watermarking (Hui et al. 2025; Ci et al. 2024a,b; Liu et al. 2024; Yang et al. 2024), traditional human-verification methods such as CAPTCHAs are becoming increasingly vulnerable, as modern deep learning and multimodal models can now break many once-secure CAPTCHA types (Bursztein, Martin, and Mitchell 2011; Shoham 1994; Wang et al. 2023). Techniques such as GANs, vision-language models (VLMs), and reinforcement learning have enabled attackers to mimic human behavior with increasing precision (Ginsberg 2012; Noury and Rezaei 2020; Schick et al. 2023). As a result, researchers have begun developing multimodal CAPTCHA datasets and evaluation frameworks to assess model performance across various CAPTCHA types (Acien et al. 2020; Farebrother, Machado, and Bowling 2019; Acien et al. 2021). This makes it vital to reassess CAPTCHA's real-world security to ensure a trustworthy Internet service (Farebrother, Machado, and Bowling 2019).

Existing studies often target specific CAPTCHA types without broad comparisons (Bursztein, Martin, and Mitchell 2011; Ci et al. 2024b; Gupta et al. 2018). The absence of a large-scale, multimodal benchmark (Hernández-Castro, Barrero, and R-Moreno 2021; Sanh et al. 2021; Ci et al. 2024a) limits systematic evaluation and hinders robust CAPTCHA design. A unified evaluation platform is urgently needed.

MCA-Bench is the first end-to-end CAPTCHA security benchmark spanning four modalities—static visual recognition, point-and-click localization, interactive manipulation and textual logic Q&A—across twenty real-world tasks. It provides over 180000 training samples and a 4000-item test set, organized into four clusters that respectively evaluate OCR robustness to visual noise, target retrieval in complex scenes, human-like interaction behaviors, and multi-step language reasoning. Representative samples from the MCA-Bench dataset are shown in Figure 1.

We use Qwen2.5-VL-7B as the vision-language backbone, fine-tuned with LoRA adapters for each task (Hu et al. 2022). Training for static and logic CAPTCHAs is supervised with target labels, while for interactive tasks, human demonstration data is used for behavior cloning. A specially designed JSON protocol facilitates large-scale evaluation and integration.

We use pass rate as the core metric. Evaluation shows multimodal VLMs exceed 96% accuracy on simple tasks but fall to as low as 2.5% on complex ones requiring physical interaction or multi-step reasoning. This reveals that visual confusion, interaction depth, and semantic complexity jointly drive attack difficulty, offering practical guidance for CAPTCHA design. MCA-Bench is open-sourced to enable reproduction and sustain iterative attack/defense benchmarking. The main contributions are as follows:

- MCA-Bench: the first large-scale, cross-modal

*Corresponding Author.

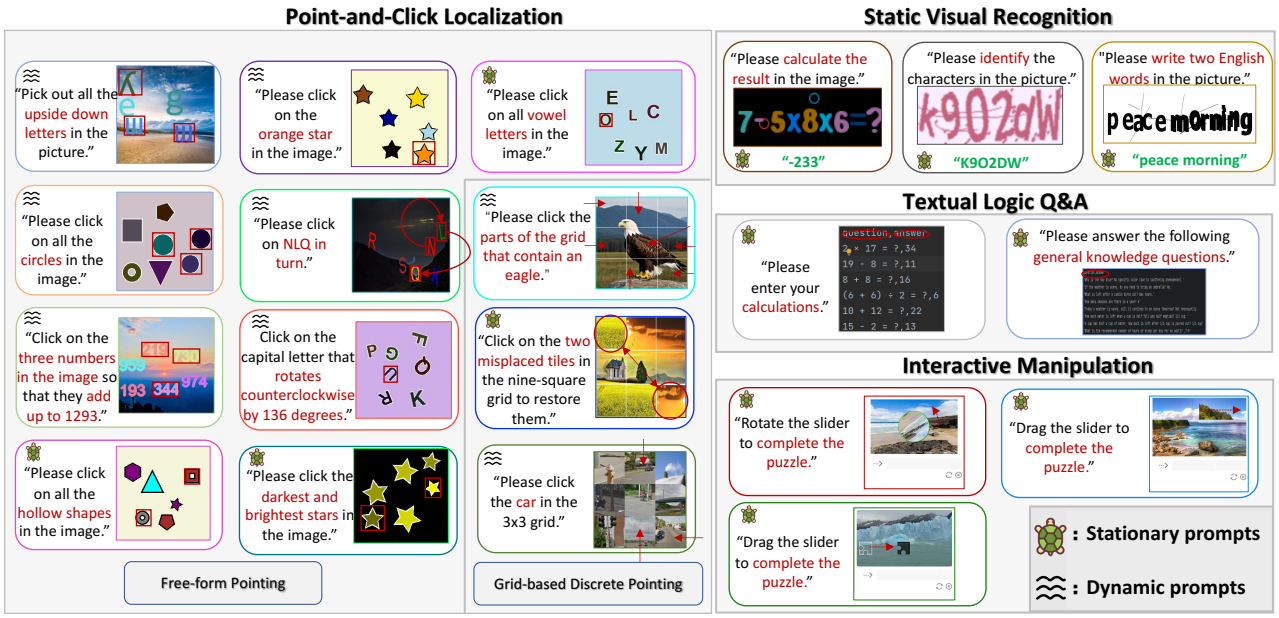

Figure 1: Data samples from MCA-Bench.Includes four categories and 20 sub-clusters of Point-and-Click Localization, Static Visual Recognition, Textual Logic Q&A and Interactive Manipulation.

CAPTCHA attack benchmark with 20 real tasks.

- Proposed a unified evaluation pipeline with a single pass-rate metric and open source scripts.
- First full-scale CAPTCHA security assessment with guidance for human–machine verification.

Related Work

VLMs in Structured Image Tasks

Visual Language Models (VLMs) unify image and text representations, with early models like CLIP (Noury and Rezaei 2020) and ALIGN (Hossen and Hei 2022) performing well on simple tasks but lacking complex reasoning. Recent approaches integrate LLMs with visual encoders to enhance cross-modal understanding—e.g., BLIP-2 (Kumar and Jindal 2021) employs frozen components for efficiency, while MiniGPT-4 (Wang et al. 2024) maps visual features to LLMs for improved QA. As focus shifts to structured inputs like tables and documents—similar to text-heavy, noisy CAPTCHA images—models such as Qwen-VL (Song et al. 2025) advance encoder design and multilingual layout understanding. Architectures like LLaVA (Li et al. 2023) and MiniGPT-4 leverage unified visual projection and instruction tuning to handle distortions and occlusions. However, most VLMs remain unadapted to CAPTCHA tasks, limiting their performance and generalization.

Evolution of Intelligent Agents Toward AGI

An AI agent is an autonomous entity that perceives, decides, and acts in its environment, characterized by autonomy, reactivity, and social interactivity—traits crucial to AGI development (Searles et al. 2023; Hong et al. 2024). Agent research has evolved from symbolic rule-based systems (Luo et al. 2024; Finn, Abbeel, and Levine 2017), limited by uncertainty and scalability (Radford et al. 2021), to reactive agents with fast perception-action loops (Russell and Wefald 1991) but limited planning. Reinforcement learning enabled agents like AlphaGo (Shah et al. 2023), though sample inefficiency and poor generalization persisted (Park et al. 2023; Jia et al. 2021). Transfer and meta-learning improved adaptation via knowledge reuse (Fakoor et al. 2020), despite high pre-training costs (Elson et al. 2007). Recently, LLM-based agents show emergent reasoning, planning (Liu et al. 2024; Achiam et al. 2023), multimodal understanding (e.g., BLIP-2 (Kaelbling, Littman, and Moore 1996)), visual generation (Song, Liu, and Shou 2025a; Huang et al. 2025; Song, Liu, and Shou 2025b), dynamic task decomposition (e.g., Voyager (Sivakorn, Polakis, and Keromytis 2016)), and tool use (e.g., Toolformer (Ribeiro 2002)), enabling general-purpose intelligence (Norvig and Russell 2021; Song et al. 2024). Their zero-shot generalization (Sumers et al. 2024) and social collaboration abilities (H. 1989) mark a paradigm shift in agent research.

Advances and Challenges in CAPTCHA Security

Early text CAPTCHAs relied on heavy distortion and noise, but CNN-based segmentation attacks soon defeated many schemes, including reCAPTCHA (Chellapilla et al. 2005; Shet 2014; Bursztein, Martin, and Mitchell 2011). Image-based CAPTCHAs, such as ASIRRA, emphasized visual cognition yet were ultimately bypassed by SVM classifiers trained on public datasets (Ding et al. 2025; Gao et al. 2021). Modern defenses now employ deep object detection, diffusion models, and style transfer for increased complexity (Mann et al. 2020; Van Le et al. 2023; Hutter 2005; Ci et al. 2024b; Ginsberg 2012; Liu et al. 2023). Interactive

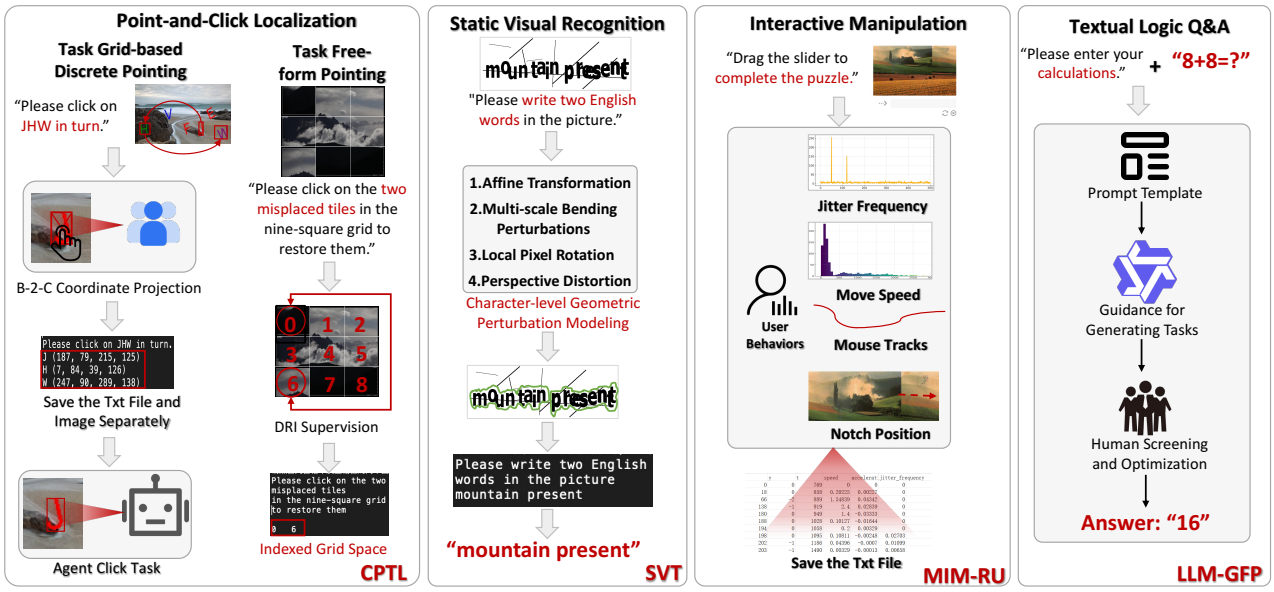

Figure 2: Schematic overview of the MCA-Bench data-acquisition and annotation workflow. From left to right, the four grey panels are Static Visual Recognition, Interactive Manipulation, Point-and-Click Localization, and Textual Logic Q&A; the red labels mark each category's data-collection pipeline. Each pipeline has four stages: (i) define the raw input format; (ii) apply task-specific geometric transforms, coordinate projections, or prompt/template generation; (iii) separate fine-grained annotation types; and (iv) save the annotations to text files.

CAPTCHAs (e.g., reCAPTCHA v2) add user gestures like clicking or dragging to resist automation (Schoppers 1987). Nonetheless, challenges remain in adaptive difficulty, device consistency, adversarial robustness, and reproducible benchmarks (Jiang et al. 2023; Alsuhibany 2016).

Dataset Construction Process

This section introduces the MCA Bench pipeline—from raw sample collection to final release—detailing processing and annotation strategies across four task clusters and twenty subtasks. See Figure 2 for the complete workflow.

Data Collection Sources and Compliance Strategies

In developing the MCA-Bench benchmark suite, we implemented a comprehensive data collection and compliance management system to uphold scientific integrity and ensure legal and ethical standards in evaluating CAPTCHA security. Our approach combines independently created data and reused public datasets. For the former, we designed a diverse CAPTCHA dataset, including distorted text, obfuscated graphics, interactive tasks, and puzzles, based on a thorough analysis of technology trends and security needs. These designs are original, validated, and proprietary to our team. For reused datasets, we carefully selected authorized academic, open-source, and industry resources, adhering to licensing agreements and privacy regulations. We ensured compliance through preprocessing steps like anonymization and mitigating privacy risks. Additionally, we established a multi-level review system with legal and ethics oversight, encrypted storage, and tiered access control to guarantee ongoing compliance and data security for MCA-Bench.

Data Collection and Processing

Text-based Task Data Collection For text-based CAPTCHA tasks, we designed a semi-automated, LLM-driven pipeline to efficiently generate and filter math and commonsense questions, minimizing manual effort while preserving quality and diversity. Using adaptive prompts, the Qwen LLM produced structured, semantically relevant QA pairs, guided by knowledge constraints, task-aware sampling, and type-controlled difficulty. Outputs were refined via manual filtering with a custom evaluation protocol assessing grammar, reasoning, and ambiguity to ensure clarity and robustness for real-world deployment.

Click-based Coordinate Task Data Collection We propose CPTL (Click-based Positioning and Target Localization), a multimodal benchmark designed to evaluate models' spatial localization and image-language alignment across varying complexities. It consists of two tasks: Free-form Pointing and Grid-based Discrete Pointing. In Free-form Pointing, we combine procedural background perturbations with public datasets, such as Flickr scenic photos, to create images rich in semantic content. The Grid-based Discrete Pointing task uses a 3×3 grid to evaluate decision-making within specific regions, dividing the image into 9 segments (0–8). The image content includes a 64-class animal dataset, Flickr scenic backgrounds, Google ReCaptcha V2 challenges, and manually designed targets.

CPTL enables adjustable evaluation of spatial complexity and instruction modalities, effectively isolating spatial parsing from language comprehension. It simulates real-world CAPTCHA conditions, testing model robustness

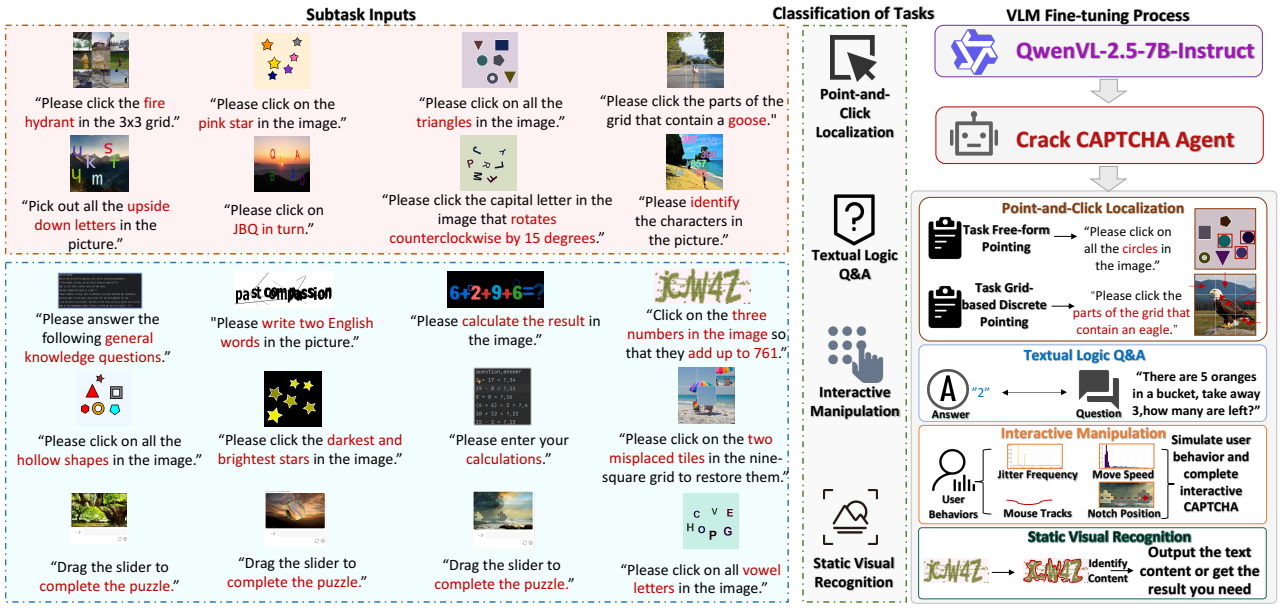

Figure 3: Schematic of Data Flow Across Four Framework Stages. The schematic diagram illustrates the data flow and key module configuration across the four stages of the end-to-end framework: unified interface access, gent fine-tuning loading, collaborative inference execution, structured result feedback.

against noisy backgrounds, multi-target interference, and weak prompts. Experiments demonstrate that models fine-tuned with CPTL achieve strong localization and semantic understanding, establishing a reliable benchmark for future CAPTCHA security research.

Static Visual Recognition Task Data Collection We propose SVT (Static Visual Textual Understanding), a benchmark to evaluate multimodal models' ability to recognize and reconstruct text from distorted images. SVT leverages procedural image generation and linguistic constraints, applying character-level geometric distortions, stochastic noise, and occlusions to create challenging yet interpretable samples. It assesses models' fine-grained visual attention and robustness to structural noise, spanning from perception to symbolic reasoning. Experiments demonstrate SVT's effectiveness in revealing limitations of pretrained multimodal models in character-level understanding.

Interactive Behavior Task Data Collection We propose a multimodal interaction modeling framework for interactive CAPTCHA tasks, leveraging real user data to enhance vision-language models' dynamic understanding. Focusing on continuous motion interactions—such as sliding alignment, rotation calibration, and trajectory restoration—the framework requires models to interpret spatial structures, motion directions, and behavioral patterns. We collected diverse user trajectories, annotated with timestamps and velocity, enabling realistic spatiotemporal modeling.

Tasks are structured around target-state restoration, divided into standardized subtasks with clear start, goal, and intermediate states. Interaction trajectories are serialized as behavioral vectors combined with visual data for joint learning, significantly enhancing action alignment. To evaluate models comprehensively, we propose new metrics: Center Deviation Error, Angular Restoration Accuracy, Slide Path Alignment Rate, and Motion Variability Index.

Data Annotation Strategy

We built a standardized, task-driven annotation framework with four representative task types to ensure consistent multimodal CAPTCHA training and evaluation.

Unified Intent Modeling for Coordinate Pointing and Grid Selection For free-form coordinate pointing tasks, we adopt a box-to-center projection strategy. Annotators label each target by marking the top-left and bottom-right corners of its bounding box using absolute pixel coordinates, with $(0, 0)$ at the image's top-left. The geometric center of the box serves as the training target. During inference, a prediction is considered correct if it falls within the box, following an IoB-Gated Validation rule. This provides spatial tolerance, improves robustness to outliers, and stabilizes training. Formally, the validation criterion for a predicted point p and bounding box $b = [\, b_{\min}, b_{\max} \,]$ is defined as:

$$\mathcal{G}(p,\, b) \;=\; \mathbb{I}\Big(\big\| \, D^{-1} \big(p - \tfrac{1}{2}(b_{\min} + b_{\max}) \big) \big\|_{\infty} \le \tfrac{1}{2} \Big) \quad (1)$$

Here, $\mathbb{I}(\cdot)$ is the indicator function (1 if the condition holds, 0 otherwise), $D = \mathrm{diag}(\, b_{\max} - b_{\min})$ is the diagonal matrix of bounding-box width and height, and $\|\cdot\|_{\infty}$ is the ℓ_{∞} norm (the maximum absolute component). Thus, the expression tests whether the predicted point lies within the normalized box region centered on the ground-truth box.

We propose a discrete region-index supervision scheme: each image is split into a 3×3 row-major grid, and the target is classified by its cell index, aligning language to local

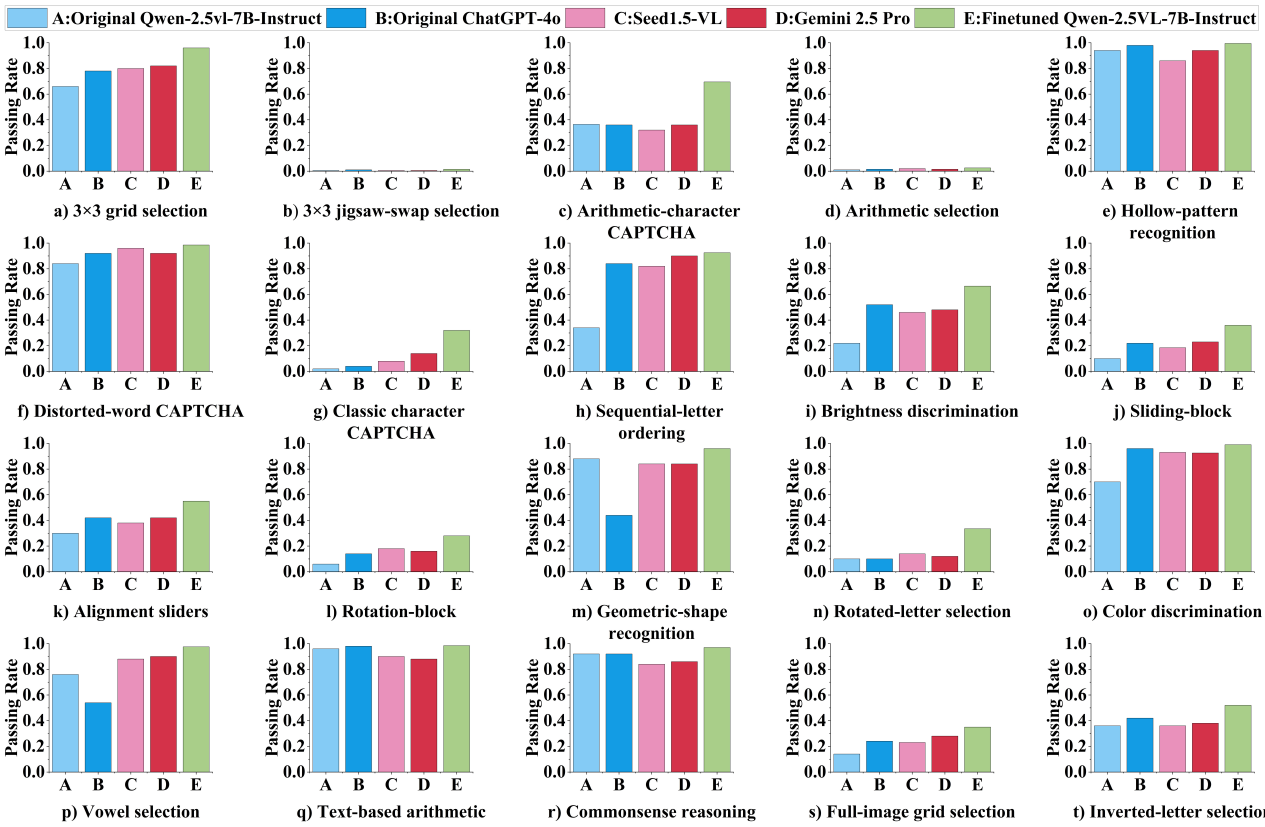

Figure 4: Performance Comparison of Multimodal Language Models on MCA-Bench CAPTCHA Tasks. The figure compares the success rates of models including Qwen2.5-VL-7B-Instruct, ChatGPT-4o, Seed1.5-VL, Gemini2.5-Pro, and fine-tuned Qwen2.5-VL-7B-Instruct across MCA-Bench CAPTCHA tasks, covering basic visual recognition, character-based recognition, and advanced multi-step reasoning challenges. Results show fine-tuning consistently improves performance, yet even top-performing models lag behind human-level robustness in complex reasoning tasks.

regions and improving robustness under multi-object interference and ambiguous or similar visual cues.

Static Visual Recognition Class Task Labeling Methods A unified labeling framework adapts to any static-vision task with task-specific supervision: for character recognition, we generate precise, content-aligned tags via automated text generation plus human review to balance speed and accuracy; for higher-level semantics, we use Visual-Symbolic Abstraction Parsing to deterministically translate visual structure into "final semantic solutions," enabling rigorous logical reasoning and computation—yielding clear, consistent labels that drive precise model training and reliable evaluation even in the most complex cases.

Interactive Behavior Task Labeling Methods We model interaction-based CAPTCHAs by capturing fine-grained user trajectories, target actions, and motion cues—whether sliding, rotating, or aligning layers. By fusing visual restoration with behavioral analysis, the system flags genuine human input (characterized by nonlinear paths, variable speed, and natural timing) and filters out the uniform, mechanically precise patterns typical of bots.

In slider CAPTCHAs, the user drags a puzzle piece to its gap; in rotation CAPTCHAs, they turn the image with a slider; in alignment CAPTCHAs, they slide layers until they coincide. Ten checkpoints record, and we score each trace for path smoothness, velocity profile, and plausible duration. For rotation, the slider's horizontal travel $\Delta x \in [0, L]$ ($L = 250$ px) maps to the counter-clockwise angle and can be computed as follows:

$$\phi(\Delta x) = \arg\left(\exp\left(i \cdot 2\pi \cdot \frac{\Delta x}{L}\right)\right) \cdot \frac{360^\circ}{2\pi} \qquad (2)$$

$\Delta x \in [0, L]$ is the slider's horizontal displacement in pixels (with $L = 250\,\text{px}$), $i^2 = -1$ the imaginary unit; $\exp\left(i \cdot 2\pi \cdot \frac{\Delta x}{L}\right)$ maps this displacement onto the complex unit circle, $\arg(\cdot)$ extracts its phase (in radians), and the factor $\frac{360^\circ}{2\pi}$ converts that phase into the rotation angle $\phi(\Delta x)$.

Agent Pipeline and Experiments

This section analyzes experimental results on MCA-Bench, comparing model performance across varying multimodal CAPTCHA difficulties. We highlight each method's strengths and limitations under different task settings and discuss implications for more robust vision–language verification systems.

Model	Pass@k	3×3 grid sel.	3×3 jig-swap	Arith. char.	Arith. sel.	Hollow pattern	Distort. word	Classic char.	Sequential letter	Bright. dist.	Sliding block
ChatGPT-4o	Pass@2	0.780	0.010	0.360	0.015	0.980	0.920	0.045	0.840	0.520	0.220
	Pass@3	0.780	0.010	0.360	0.020	0.985	0.920	0.050	0.840	0.520	0.225
	Pass@4	0.780	0.010	0.360	0.020	0.985	0.920	0.050	0.840	0.520	0.225
	Pass@5	0.780	0.015	0.360	0.020	0.990	0.920	0.050	0.840	0.520	0.225
Seed1.5-VL	Pass@2	0.800	0.005	0.320	0.020	0.865	0.960	0.085	0.820	0.465	0.185
	Pass@3	0.800	0.005	0.320	0.020	0.865	0.960	0.085	0.825	0.465	0.190
	Pass@4	0.805	0.005	0.320	0.020	0.870	0.960	0.085	0.825	0.465	0.195
	Pass@5	0.805	0.005	0.320	0.020	0.875	0.960	0.085	0.825	0.465	0.195
Gemini2.5-Pro	Pass@2	0.820	0.005	0.360	0.015	0.940	0.920	0.140	0.905	0.480	0.230
	Pass@3	0.825	0.005	0.360	0.015	0.945	0.920	0.140	0.905	0.480	0.235
	Pass@4	0.825	0.010	0.360	0.020	0.945	0.920	0.145	0.910	0.485	0.235
	Pass@5	0.825	0.010	0.360	0.020	0.945	0.920	0.150	0.910	0.485	0.235
Original Qwen-2.5vl-7B-Instr.	Pass@2	0.660	0.010	0.365	0.010	0.940	0.020	0.025	0.340	0.220	0.100
	Pass@3	0.660	0.010	0.365	0.010	0.940	0.020	0.025	0.340	0.220	0.100
	Pass@4	0.665	0.015	0.365	0.010	0.950	0.025	0.030	0.345	0.225	0.100
	Pass@5	0.665	0.015	0.365	0.010	0.950	0.025	0.030	0.345	0.225	0.100
Human	Pass@2	0.960	0.980	0.990	0.955	0.995	0.885	0.930	0.960	0.965	0.980
	Pass@3	0.985	0.985	0.990	0.955	0.995	0.965	0.950	0.980	0.985	0.985
	Pass@4	0.995	0.985	0.995	0.965	1.000	0.985	0.980	0.985	0.985	0.990
	Pass@5	1.000	0.990	1.000	0.970	1.000	0.990	0.995	0.995	0.990	0.995
Finetuned Qwen-2.5vl-7B-Instr.	Pass@2	0.965	0.015	0.700	0.025	0.995	0.985	0.325	0.925	0.665	0.365
	Pass@3	0.970	0.015	0.705	0.025	0.995	0.985	0.325	0.925	0.665	0.365
	Pass@4	0.975	0.015	0.705	0.030	0.995	0.985	0.325	0.925	0.670	0.365
	Pass@5	0.980	0.020	0.710	0.035	1.000	0.995	0.330	0.930	0.675	0.365
Model	Pass@k	Align. sliders	Rotate block	Geom. shape	Rotat. letter	Color discr.	Vowel sel.	Full-img grid sel.	Text-based arith.	Common sense	Invert. letter
ChatGPT-4o	Pass@2	0.420	0.140	0.440	0.100	0.960	0.540	0.240	0.980	0.925	0.425
	Pass@3	0.420	0.140	0.440	0.100	0.960	0.540	0.240	0.985	0.925	0.425
	Pass@4	0.420	0.140	0.440	0.100	0.960	0.540	0.245	0.985	0.925	0.430
	Pass@5	0.420	0.140	0.440	0.100	0.960	0.540	0.245	0.985	0.930	0.430
Seed1.5-VL	Pass@2	0.380	0.180	0.840	0.140	0.930	0.880	0.230	0.905	0.840	0.360
	Pass@3	0.380	0.180	0.840	0.140	0.930	0.880	0.235	0.910	0.845	0.365
	Pass@4	0.380	0.185	0.840	0.140	0.930	0.880	0.235	0.920	0.845	0.365
	Pass@5	0.380	0.185	0.845	0.140	0.935	0.880	0.235	0.920	0.850	0.365
Gemini2.5-Pro	Pass@2	0.420	0.160	0.845	0.120	0.925	0.900	0.280	0.880	0.860	0.385
	Pass@3	0.420	0.160	0.845	0.120	0.925	0.900	0.280	0.885	0.865	0.385
	Pass@4	0.420	0.165	0.845	0.120	0.925	0.905	0.285	0.890	0.870	0.390
	Pass@5	0.420	0.165	0.850	0.125	0.930	0.905	0.290	0.895	0.875	0.400
Original Qwen-2.5vl-7B-Instr.	Pass@2	0.300	0.060	0.885	0.100	0.700	0.760	0.140	0.960	0.920	0.365
	Pass@3	0.305	0.060	0.885	0.100	0.700	0.760	0.140	0.960	0.920	0.365
	Pass@4	0.305	0.060	0.885	0.100	0.700	0.760	0.140	0.965	0.925	0.365
	Pass@5	0.305	0.065	0.885	0.100	0.700	0.760	0.145	0.965	0.925	0.370
Human	Pass@2	0.975	0.965	0.990	0.875	0.970	0.960	0.940	0.970	0.875	0.955
	Pass@3	0.995	0.990	0.990	0.950	0.975	0.980	0.975	0.985	0.905	0.985
	Pass@4	1.000	0.995	0.995	0.985	0.990	0.995	0.990	0.990	0.955	1.000
	Pass@5	1.000	0.995	1.000	0.990	0.995	1.000	0.995	1.000	0.985	1.000
Finetuned Qwen-2.5vl-7B-Instr.	Pass@2	0.565	0.285	0.960	0.335	0.990	0.975	0.360	0.985	0.975	0.520
	Pass@3	0.565	0.285	0.960	0.335	0.990	0.975	0.360	0.985	0.975	0.520
	Pass@4	0.565	0.285	0.960	0.335	0.995	0.975	0.370	0.990	0.980	0.525
	Pass@5	0.570	0.285	0.960	0.340	0.995	0.980	0.370	0.995	0.990	0.530

Table 1: Performance of vision-language models on CAPTCHA-Bench. This table presents Pass@k accuracy on 20 CAPTCHA task types, comparing models and human performance.

Set Up

We fine-tune LoRA adapters on Qwen2.5-VL-7B-Instruct across 20 CAPTCHA tasks (Hu et al. 2022). Inputs are 224×224 images paired with structured prompts. Training runs on 4 H20 GPUs (batch size 8, gradient accumulation 4 $\rightarrow$ effective 32) using AdamW with a linear LR decay from from 1×10^{-4}. We save checkpoints every 100 steps and stop early if validation loss doesn't improve over 20 evaluations.

Dataset	Qwen2.5-VL-7B	Human
Point-and-click localization		
3×3 grid selection	0.960	0.880
Inverted-letter selection	0.520	0.940
Geometric-shape recognition	0.960	0.980
Brightness discrimination	0.665	0.780
Hollow-pattern recognition	0.995	0.985
Sequential-letter ordering	0.925	0.980
Full-image grid selection	0.350	0.740
Color discrimination	0.990	0.885
Vowel selection	0.975	0.805
Arithmetic selection	0.025	0.780
Rotated-letter selection	0.335	0.745
3×3 jigsaw-swap selection	0.015	0.805
Static visual recognition		
Classic character CAPTCHA	0.320	0.920
Distorted-word CAPTCHA	0.985	0.840
Arithmetic-character CAPTCHA	0.695	0.985
Interactive manipulation		
Sliding-block	0.360	0.740
Rotation-block	0.280	0.760
Alignment sliders	0.550	0.720
Textual logic Q&A		
Text-based arithmetic	0.985	0.970
Commonsense reasoning	0.970	0.860

Table 2: Comparison of Pass Rates for CAPTCHA Types: Qwen2.5-VL-7B vs. Human Performance.

MCA-Bench: Cracking Capability Dataset

Fine-tuning QWen-2.5VL-7B-instruct on MCA-Bench yields substantial gains. As shown in Fig 4, the adapted model surpasses its zero-shot baseline and closed-source peers across CAPTCHA tasks, enhancing visual recognition, logical reasoning, interaction, and robustness to complex challenges.

MCA-Bench Dataset Overview

MCA-Bench is the first multimodal CAPTCHA dataset covering visual recognition, point selection, textual reasoning, and interactive operations. It features varied image sizes, publicly sourced tasks for diversity and reproducibility, and a large training set that supports multi-image fine-tuning to improve generalization.

Agent System Pipeline

The pipeline packs every image, prompt, and user action into one JSON record. A task-ID in the header activates the matching LoRA agent (Hu et al. 2022), which adds its lightweight adapter to a shared frozen backbone. Visual- and text-embeddings flow through that agent to generate outputs—coordinates, strings, mouse traces, or character codes. These results are written back into the original JSON format for downstream use. This design lets the same backbone handle all CAPTCHA tasks with minimal memory and no intermediate parsers (Fig. 3).

Fine-Tuned Model Performance in Visual Tasks

On simple visual tasks with minimal noise, the fine-tuned VLM surpasses human speed and accuracy. However, its performance sharply declines under complex transformations such as distortion or rotation, highlighting VLMs' limitations compared to robust human perception.

Zero-Shot vs. Fine-Tuned Performance Gaps

Despite differing architectures and training regimes, zero-shot pass rates vary widely across Qwen2.5-VL, ChatGPT-4o, Seed1.5-VL, and Gemini2.5-Pro (Table 1), yet none match the fluid adaptability of human solvers. Fine-tuning on MCA-Bench consistently improves performance—especially for Qwen2.5-VL—but even the best-tuned agents remain below human robustness on multi-step reasoning or precise interaction tasks. This highlights that, while instruction tuning and backbone advances bring notable gains, human proficiency in nuanced, context-rich challenges is still unmatched.

Human vs. AI in Reasoning and Interaction

In tasks requiring reasoning or behavioral interaction—such as sliding puzzles, rotating tiles, multi-step reasoning, and common-sense judgment—the fine-tuned model still lags behind human performance. As shown in Table 2, despite improvements in standard visual recognition, the fine-tuned Qwen-Agent struggles with tasks demanding deeper understanding, contextual reasoning, and precise coordination.

Conclusion

MCA-Bench is the first unified benchmark showing that although LoRA-based attack agents achieve over 96% accuracy on visual and shallow-semantic CAPTCHAs, their performance drops below 2.5% on tasks involving physical interaction or multi-step reasoning—highlighting the limits of single-dimensional obfuscation. To address this gap, we introduce a tightly integrated human–machine verification paradigm built on three principles: (1) deep modality coupling — unifying visual cue localization, logical reasoning, and interactive input; (2) behavior-anchored validation — leveraging human interaction trajectories such as timing and continuity; and (3) session-specific semantic personalization — injecting unique semantic context into each challenge. Together, these principles form a cognitively rich, adaptive verification framework that significantly enhances robustness against advanced AI-driven attacks.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Acien, A.; Morales, A.; Fierrez, J.; Vera-Rodriguez, R.; and Bartolome, I. 2020. BeCAPTCHA: Detecting human behavior in smartphone interaction using multiple inbuilt sensors. *arXiv preprint arXiv:2002.00918*.

Acien, A.; Morales, A.; Fierrez, J.; Vera-Rodriguez, R.; and Delgado-Mohatar, O. 2021. BeCAPTCHA: Behavioral bot detection using touchscreen and mobile sensors benchmarked on HuMIdb. *Engineering Applications of Artificial Intelligence*, 98: 104058.

Alsuhibany, S. A. 2016. A benchmark for designing usable and secure text-based captchas. *International Journal of Network Security & Its Applications*, 8(4): 41–54.

Bursztein, E.; Martin, M.; and Mitchell, J. 2011. Text-based CAPTCHA strengths and weaknesses. In *Proceedings of the 18th ACM conference on Computer and communications security*, 125–138.

Chellapilla, K.; Larson, K.; Simard, P. Y.; and Czerwinski, M. 2005. Building segmentation based human-friendly human interaction proofs (HIPs). In *International Workshop on Human Interactive Proofs*, 1–26. Springer.

Ci, H.; Song, Y.; Yang, P.; Xie, J.; and Shou, M. Z. 2024a. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*.

Ci, H.; Yang, P.; Song, Y.; and Shou, M. Z. 2024b. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*, 338–354. Springer.

Ding, Z.; Deng, G.; Liu, Y.; Ding, J.; Chen, J.; Sui, Y.; and Li, Y. 2025. IllusionCAPTCHA: A CAPTCHA based on visual illusion. In *Proceedings of the ACM on Web Conference 2025*, 3683–3691.

Elson, J.; Douceur, J. R.; Howell, J.; and Saul, J. 2007. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. *CCS*, 7(366-374): 15.

Fakoor, R.; Chaudhari, P.; Soatto, S.; and Smola, A. 2020. Meta-Q-Learning.

Farebrother, J.; Machado, M. C.; and Bowling, M. 2019. Generalization and Regularization in DQN.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1126–1135. JMLR.org.

Gao, Y.; Gao, H.; Luo, S.; Zi, Y.; Zhang, S.; Mao, W.; Wang, P.; Shen, Y.; and Yan, J. 2021. Research on the security of visual reasoning {CAPTCHA}. In *30th USENIX security symposium (USENIX security 21)*, 3291–3308.

Ginsberg, M. 2012. *Essentials of Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 9780323139687.

Gupta, A.; Mendonca, R.; Liu, Y.; Abbeel, P.; and Levine, S. 2018. Meta-reinforcement learning of structured exploration strategies. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 5307–5316. Red Hook, NY, USA: Curran Associates Inc.

H., W. C. J. C. 1989. Learning With Delayed Rewards. *Ph. D. thesis, Cambridge University*.

Hernández-Castro, C. J.; Barrero, D. F.; and R-Moreno, M. D. 2021. BASECASS: A methodology for CAPTCHAs security assurance. *Journal of Information Security and Applications*, 63: 103018.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Zhang, C.; Wang, J.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; Xiao, L.; Wu, C.; and Schmidhuber, J. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv:2308.00352.

Hossen, I.; and Hei, X. 2022. aaecaptcha: The design and implementation of audio adversarial captcha. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 430–447. IEEE.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, S.; Song, Y.; Zhang, Y.; Guo, H.; Wang, X.; Shou, M. Z.; and Liu, J. 2025. PhotoDoodle: Learning Artistic Image Editing from Few-Shot Pairwise Data. arXiv:2502.14397.

Hui, S.; Song, Y.; Zhou, S.; Deng, Y.; Huang, W.; and Wang, J. 2025. Autoregressive Images Watermarking through Lexical Biasing: An Approach Resistant to Regeneration Attack. arXiv:2506.01011.

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*, volume 1. Springer, 1st edition. ISBN 3540221395.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 4904–4916. PMLR.

Jiang, R.; Zhang, S.; Liu, L.; and Peng, Y. 2023. Diff-CAPTCHA: An Image-based CAPTCHA with Security Enhanced by Denoising Diffusion Model. *arXiv preprint arXiv:2308.08367*.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement Learning: A Survey. *arXiv e-prints*, cs/9605103.

Ke, Z.; Cao, Y.; Chen, Z.; Yin, Y.; He, S.; and Cheng, Y. 2025. Early warning of cryptocurrency reversal risks via multi-source data. *Finance Research Letters*, 107890.

Kumar, M.; and Jindal, M. K. 2021. Benchmarks for designing a secure devanagari CAPTCHA. *SN Computer Science*, 2: 1–16.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image

encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34892–34916. Curran Associates, Inc.

Liu, Y.; Song, Y.; Ci, H.; Zhang, Y.; Wang, H.; Shou, M. Z.; and Bu, Y. 2024. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*.

Luo, H.; Gu, J.; Liu, F.; and Torr, P. 2024. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv preprint arXiv:2403.09766*.

Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1: 3.

Norvig, P.; and Russell, S. 2021. *Artificial Intelligence: A Modern Approach*. USA: Pearson Education.

Noury, Z.; and Rezaei, M. 2020. Deep-CAPTCHA: a deep learning based CAPTCHA solver for vulnerability assessment. *arXiv preprint arXiv:2006.08296*.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ribeiro, C. 2002. Reinforcement Learning Agents. *Artif. Intell. Rev.*, 17(3): 223–250.

Russell, S.; and Wefald, E. 1991. *Do the right thing: studies in limited rationality*. Cambridge, MA, USA: MIT Press. ISBN 0262181444.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.

Schoppers, M. J. 1987. Universal plans for reactive robots in unpredictable environments. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'87, 1039–1046. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Searles, A.; Nakatsuka, Y.; Ozturk, E.; Paverd, A.; Tsudik, G.; and Enkoji, A. 2023. An empirical study & evaluation of modern {CAPTCHAs}. In *32nd usenix security symposium (usenix security 23)*, 3081–3097.

Shah, D.; Osiński, B.; Levine, S.; et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, 492–504. PMLR.

Shet, V. 2014. Are you a robot? introducing no captcha recaptcha. *Google Security Blog*, 3: 12.

Shoham, Y. 1994. Agent oriented programming: An overview of the framework and summary of recent research. In Masuch, M.; and Pólos, L., eds., *Knowledge Representation and Reasoning Under Uncertainty*, 123–129. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-48451-6.

Sivakorn, S.; Polakis, I.; and Keromytis, A. D. 2016. I am robot:(deep) learning to break semantic image captchas. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 388–403. IEEE.

Song, Y.; Liu, C.; and Shou, M. Z. 2025a. MakeAnything: Harnessing Diffusion Transformers for Multi-Domain Procedural Sequence Generation. arXiv:2502.01572.

Song, Y.; Liu, C.; and Shou, M. Z. 2025b. OmniConsistency: Learning Style-Agnostic Consistency from Paired Stylization Data. arXiv:2505.18445.

Song, Y.; Lou, S.; Liu, X.; Ci, H.; Yang, P.; Liu, J.; and Shou, M. Z. 2024. Anti-Reference: Universal and Immediate Defense Against Reference-Based Generation. *arXiv preprint arXiv:2412.05980*.

Song, Y.; Yang, P.; Ci, H.; and Shou, M. Z. 2025. Idprotector: An adversarial noise encoder to protect against id-preserving image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3019–3028.

Sumers, T. R.; Yao, S.; Narasimhan, K.; and Griffiths, T. L. 2024. Cognitive Architectures for Language Agents. arXiv:2309.02427.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2116–2127.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and Wen, J. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Yang, P.; Ci, H.; Song, Y.; and Shou, M. Z. 2024. Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems*, 37: 56644–56673.

Zhang, Z.; Shen, Q.; Hu, Z.; Liu, Q.; and Shen, H. 2025. Credit risk analysis for SMEs using graph neural networks in supply chain. In *Proceedings of the 2025 International Conference on Big Data, Artificial Intelligence and Digital Economy*, 81–85.