BEYOND PIXELS: BRIDGING THE COGNITIVE GAP IN AI-GENERATED VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in video generation have produced highly realistic and temporally coherent videos that are increasingly difficult to distinguish from real content. While beneficial for many applications, such progress also raises concerns such as misinformation. Existing detection methods mainly rely on pixel level artifacts and shallow spatiotemporal inconsistencies, which often fail as generation quality improves. Leveraging the cognitive process, humans can keenly detect the unrealistic aspects in AI-generated videos of high quality. Inspired by this, we propose a framework that utilizes deeper cognitive understanding beyond low level signals to detect AI-generated videos. This framework defines five cognitive scenarios, each containing different levels of cognitive information, ranging from perception and physical plausibility to semantic, logical, and social reasoning. Through a study on how human and current MLLM detect AI-generated videos, we show detection performance differences across various cognitive scenarios, which are related to the quality of the AI-generated video and the video understanding abilities of humans and models. Building on these findings and our proposed cognitive framework, we introduce an Iterative Cognitive Layered Instruction Evolution framework that enables multimodal models to evolve detection instructions under structured cognitive guidance. Experiments demonstrate improved detection performance and strong generalization across models and unseen AI-generated videos.

1 INTRODUCTION

In recent years, the field of video generation has made significant advancements. Current video generation models have demonstrated the ability to produce temporally coherent and visually impressive videos, showcasing broad applicability across multiple domainsFoo et al. (2025), including film productionUddin et al. (2025), entertainmentSchatten (2024), and advertisingLiu & Yu (2023). These high-quality AI-generated videos have reached a level of similarity to real videos, making it difficult for humans to distinguish them based on pixel quality. This raises challenges such as the spread of misinformation and content authenticity verification, posing risks to individual privacy and national security Kaur et al. (2024). To address these issues, several methods have been explored for detecting AI-generated videosMa et al. (2024); Bai et al. (2024); He et al. (2024), but previous detection methods mainly focus on the pixel level artifacts, and rely on extracting shallow spatiotemporal inconsistencies, noise patterns, or optical flow anomalies. These methods typically require large-scale data for training, and as the generation quality improves and the video generation models continue to evolve rapidly, existing methods struggle to achieve good detection performance.

Although AI-generated videos have very high quality, humans can still keenly detect the unrealistic aspects in them. This is because humans not only rely on the pixel level artifacts but also possess deeper cognitive knowledge, including logical reasoning of events and physical laws reflected in the videos. Thus we can use deeper cognitive knowledge to detect AI-generated videos of high quality. Currently, advanced multimodal models can effectively understand videos Wang et al. (2025c); Liu et al. (2025a); Wang et al. (2025f), making them strong candidates for detectors, especially since they can work in a zero-shot manner Liu et al. (2025b); Wang et al. (2025b). Building on this, we initially explore the effectiveness of using these multimodal models for video detection and find that many factors influence the final detection results, with sensitivity to prompts being the most significant, which is not surprising given the nature of these models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To better guide multimodal models in detecting AI-generated videos, we first investigate how humans distinguish real videos from AI-generated ones. Drawing from visual cognition theory, which posits that human video understanding involves complex interactions ranging from low-level visual perception to high-level semantic reasoningCavanagh (2011). Based on this, we propose a framework that defines five cognitive scenarios. These include **Perception and Senses Scenario**, **Physical Plausibility Scenario**, **Semantics and Narrative Scenario**, **Reasoning and Logic Scenario**, and **Social and Cultural Scenario**. Each scenario focuses on a distinct aspect of cognitive information. We conduct a subjective experiment on videos, where participants are asked to determine whether each video is AI-generated or real. We also evaluate different multimodal models under the same settings. Our study shows that cognitive information-based detection guidance alters the judgment logic of both humans and models. It shifts them from relying on intuition or superficial features to an understanding and analysis based on logic and content. The impact is mainly related to the quality of the AI-generated video and the ability to understand the video.

Through comparative analysis of detection results from multimodal models and humans, we observe the following key findings: ***Detecting AI-generated videos requires a flexible cognitive guidance strategy, as different video qualities and detection models exhibit significant differences in their sensitivity to cognitive information.*** Therefore, constructing detection instructions that effectively integrate information across multiple cognitive levels is essential. Manual prompt engineering has the drawbacks of relying on human expertise and the time-consuming nature of manual testingGu et al. (2023), limiting the generalizability and efficiency for detection. Building on this insight, we propose an Iterative Cognitive-Layered Instruction Evolution framework. Our framework is an instance-driven instruction self-evolution framework for AI-generated video detection. Instead of relying on manually designed prompts, our approach enables a multimodal model to iteratively refine its own detection instruction by reflecting on its understanding of AI-generated videos under structured cognitive guidance. Based on the cognitive framework we constructed, this approach incorporates cognitive guidance into the detection process, improving the ability of MLLMs to interpret video content and detect videos, thus avoiding prompt engineering. Experiments show that our evolution method can achieve good detection performance, work with different detection models and generalize to unseen AI-generated videos.

- To better detect, we introduce the CogDetect framework based on fundamental human cognitive processes, categorized by distinct cognitive reasoning levels.
- We are the first comprehensive study comparing human vs. MLLM detection performance in video generation through progressive cognitive hierarchy.
- To mirror the human cognitive process for video understanding, we propose an evolutionary detection framework, extensive experiments demonstrate that our self-evolution framework significantly enhances detection performance.

2 Related Works

AI-generated Video Detection This rapid development of generative models, such as OpenSora Zheng et al. (2024), VideoCrafter Chen et al. (2024a), and Sora OpenAI (2024), blurs the line between AI-generated and real videos, which has increased the risks associated with misinformation and authenticity verification Kaur et al. (2024). As a result, there is an urgent demand for detecting AI-generated videos. Early research focuses mainly on facial forgery Dolhansky et al. (2020); Li et al. (2020). Recent methods start to identify AI-generated videos at the pixel level beyond face. Notable methods focus on identifying optical flow anomalies Bai et al. (2024), or extracting shallow spatiotemporal inconsistencies Ma et al. (2024) He et al. (2024) Chen et al. (2024b). Some methods also go beyond pixel level analysis, such as the NSG-VD method Zhang et al. (2025a) achieves excellent detection performance based on physical continuity. Furthermore, Fu et al. (2025) identify AI-generated videos based on human-perceived fakeness. Despite these advances, existing methods are insufficient to comprehensively capture human visual cognition on AI-generated videos. Additionally, these methods often require large-scale data for training, making it challenging to keep up with the rapid evolution of generative AI. At the same time, within the training-free paradigm, models can achieve scalable and efficient effects in practical applications through prompt optimization and self-improvement mechanismsTian et al. (2024). For example, PhyT2V Xue et al. (2025) refines text prompts without involving any model retraining efforts, expanding the current T2V model's capability of video generation. In this context, we introduce an instruction self-evolution framework

that mirrors the human cognitive process for video understanding, enabling MLLMs to detect AI-generated videos in a zero-shot setting.

MLLMs for Video Understanding MLLMs Wang et al. (2024); Chen et al. (2024c) enable sophisticated video reasoning due to their strong visual perception and understanding capabilities. Correspondingly, the evaluation benchmark for the video understanding capabilities of MLLMs has evolved from early frame-level quality Han et al. (2025); Liu et al. (2024) and semantic alignment Zheng et al. (2025) assessments to higher-level evaluations, encompassing video quality understanding Wang et al. (2025c)Zhang et al. (2025b), long-form video understanding Wang et al. (2025e), and 3D spatial and temporal understanding Wang et al. (2025d). Further assessments address world modeling Liu et al. (2025a), temporal causality Wang et al. (2025a), and embodied reasoning capabilities Wang et al. (2025f). Moreover, UVE-Bench Liu et al. (2025b) leverages MLLMs to evaluate AI-generated videos across four primary dimensions, Static Quality, Temporal Quality, Dynamic Degree, and Video-Text Alignment. Wang et al. (2025b) reveals that advanced generative models can fool even expert VLM reviewers. These works demonstrate that, despite some limitations, MLLMs have made significant progress in capturing high-level semantic, temporal, and causal cues in videos, providing a solid foundation for understanding and detecting AI-generated content.

3 Five-Level Cognitive Scenarios Design

Human visual cognition is the process of interpreting and inferring visual scenes using partial information Cao et al. (2024), encompassing a hierarchical cognitive process from low-level perception to high-level reasoning Summerfield & Egner (2009); Cavanagh (2011); Alvarez (2011); Humphreys & Bruce (2021); Petrusel et al. (2017). To better utilize cognitive information for detecting AI-generated video, we constructed progressive cognitive scenes that include different levels of cognitive information, ranging from low-level visual perception to high-level reasoning and social logic. The cognitive information and video frame sequence visualizations for each cognitive scenario are shown in Figure 1, with more visualizations in Appendix A. To better understand the selected cognitive scenarios, we further conducted experiments to explore the guiding role of mixed scenarios in Appendix G.

Zero-Knowledge Scenario: This scenario serves as the baseline. Participants, including both humans and models, are tasked with detection relying entirely on raw intuition or pre-trained latent features. They decide directly whether the video is AI-generated, without cognitive information guiding them.

Perception and Senses Scenario: At this level, we emphasize visual quality, color consistency, and structural integrity of 3D depth, investigating whether models can simulate human sensitivity to frame exception. Assuming the subjects are in a purely perceptual environment, such as appreciating a high-quality anime, they should focus on single-image visual quality, lighting consistency, color transitions, smoothness of two-dimensional edges, and three-dimensional spatial alignment and depth coherence, being sensitive to samples that violate color optical principles and spatial layout rules.

Physical Plausibility Scenario: At this level, the focus is on analyzing the consistency of physical phenomena in the video with those in the real world, and exploring whether the model can understand the physical laws of the real world. Assuming the subjects are observing a scene containing multiple physical interactions, such as a car race or an object falling experiment, they will pay attention to the realism of motion, the stability of objects, their interactions with each other, and the representation of material properties. Any phenomenon that violates the laws of physics, such as abnormal deformation of human actions or unreasonable motion trajectories, will raise doubts about the video's authenticity.

Semantics and Narrative Scenario: At this level, participants will assess whether the narrative structure in the video is reasonable, exploring if the model can understand and accurately construct narratives that align with real-world logic. Assuming the subjects are watching a short film or a story segment, they will pay attention to the temporal coherence, narrative consistency, consistency between scenes and backgrounds, environmental interactions, and camera control. Any unreasonable jumps in the timeline or plot development may disrupt the narrative flow.

Reasoning and Logic Scenario: Participants need to use logical reasoning to analyze the rationality of video content, exploring whether the model can understand and apply logical reasoning to reasonably infer causal relationships and character actions. Assuming the subjects watch a complex video involving multiple character interactions and event developments, they will judge whether the characters' actions are logical, consistent with their character settings, emotionally coherent, and free from common-sense contradictions or abrupt plot changes. Any reasoning that violates common sense or sudden changes in character behavior will affect the judgment of the video's rationality.

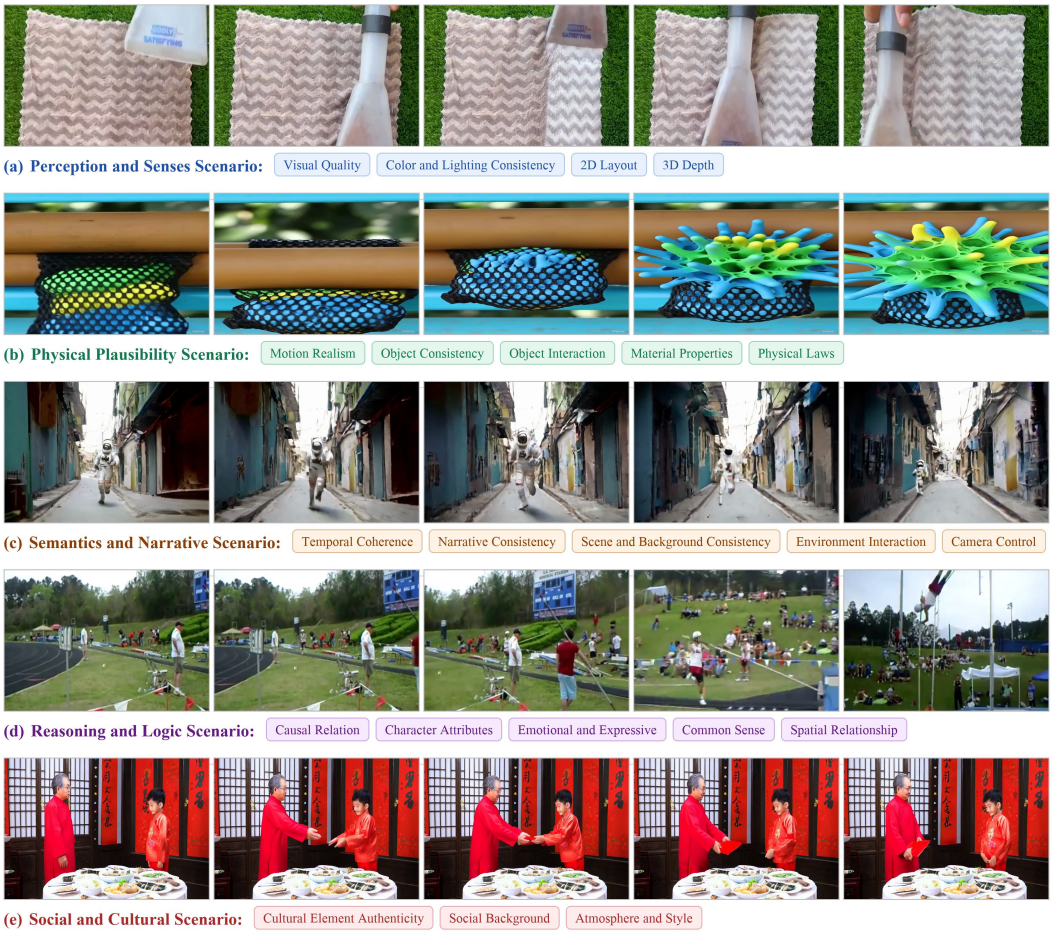

Figure 1: **Cognitive information and Visualization of Five Cognitive Scenarios Based on Human Video Understanding.** These five scenarios focusing on a distinct aspect of cognitive information, with corresponding video frame sequences visualizations for each cognitive scenario. Here, (a) and (d) represent real videos, while (b), (c), and (e) are AI-generated videos.

Social and Cultural Scenario: In this scenario, participants evaluate whether the social and cultural context presented in the video aligns with the cultural understanding and social norms of the target audience, exploring if the model can comprehend cultural backgrounds, social customs, and behavioral standards. Assuming the subjects watch a video reflecting a specific cultural context, such as a festival celebration or social event in a particular region, they will assess the authenticity of cultural elements, the social background and atmosphere, and the overall style's consistency. Any details that do not conform to local culture or social expectations, such as unnatural cultural elements or inappropriate behaviors for the context, will raise doubts about the video's cultural authenticity.

4 Human vs. MLLM in AI-generated Video Detection

We explore the use of cognitive knowledge to enhance the detection of AI-generated videos. Given the cognitive scenario constructed in Sec. 3 , an interesting question arises: ***How does different cognitive information-based detection guidance affect model and human detection performance?*** To answer this question, we compare the changes in detection performance of humans and different models across various cognitive scenarios. The experiment uses ***S**uccess **R**ate (%), **F**alse **P**ositive **R**ate (%), **F**alse **N**egative **R**ate (%), **R**ecall (%), and **F1**-score (%)* as evaluation metrics. Finally, in order to better guide MLLMs in detecting AI-generated videos, we also studied ***What factors affect the model's detection performance under different cognitive scenarios?*** Our results are shown in Figure 2, more detailed data and visualizations can be found in Appendix B .

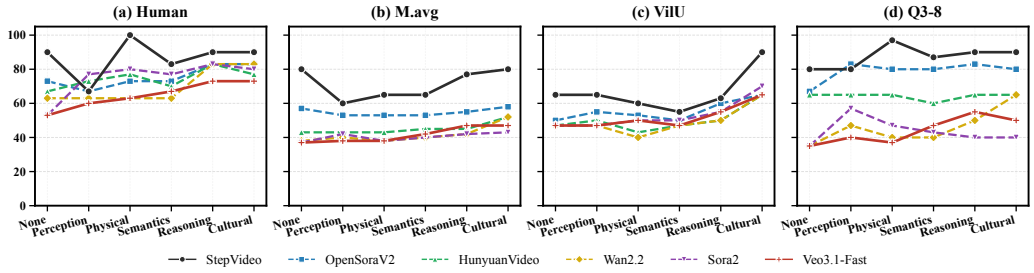

Figure 2: ***Success Rate (SR)* Comparison between Human and MLLM Detection across Various Scenarios.** The results use *SR* as the metric to demonstrate the detection performance of humans and MLLMs under cognitive information-based detection guidance. Specifically, we present (a) human average performance and (b) MLLMs average performance, and two representative models: (c) Vila-U-7b and (d) Qwen3-VL-8B. More detailed visualization results are provided in Appendix B.

CogDetect Dataset Construction To construct the CogDetect dataset, we selected 2411 videos from VideoRealityTest Wang et al. (2025b), UVE-Bench Liu et al. (2025b), Kinetics-400 Kay et al. (2017), and DeeptraceReward Fu et al. (2025). For the convenience of human evaluation, we selected videos generated by six generative models and real videos. The selected videos are further divided into six subsets based on six generative models, and we finally selected a total of 341 videos for human evaluation. Following the performance of video generation models presented in Wang et al. (2025b), we categorize the selected models into three types. These include low-quality generation models *StepVideo* with lower performance, medium-quality generation models *Opensora-V2*, *Sora2*, *HunyuanVideo* and *Wan2.2*, and high-quality generation models *Veo3.1-Fast* with higher performance. To further analyze the factors affecting model-based detection performance across various cognitive scenarios, we randomly sample videos, covering both real videos and AI-generated videos. More details about the CogDetect dataset can be found in Appendix B.1.

4.1 How Human Performs in Detection

Human Study Setting We recruit 24 human evaluators from diverse backgrounds to assess the authenticity of videos across various scenarios. We ask evaluators to judge whether each video is real or AI-generated, where the ratio of real to AI-generated videos is fixed at 1:2 in each scenario. Furthermore, to gain insights into human judgments of video authenticity, evaluators are provided with examples and open-ended questions. The results are shown in Figure 2 (a), while details of the evaluators and statistical analyses of human performance are provided in Appendix C.

Influence of Cognitive Scenario We find that cognitive information-based guidance effectively influences detection performance. The cognitive guidance shifts humans from intuitive judgment based on experience to more detailed visual and perceptual analysis. As cognitive information progresses gradually, humans' focus shifts correspondingly when understanding the video. From the **Zero-Knowledge Scenario**, where no detection guidance is provided, to the lower-level **Perception and Senses Scenario**, and then to the higher-level **Social and Cultural Scenario**, human's understanding of the video gradually deepens, the *SR* of video authenticity detection is also gradually improving. For example, under lower-level cognitive scenarios, humans focus on the video at the pixel level, such as single-image visual quality. In higher-level cognitive scenarios, humans develop a deeper understanding of the video, considering rationality or social and cultural context. This phenomenon aligns with human cognitive habits and further supports the rationality and meaningfulness of the hierarchical cognitive scenarios we have constructed.

Noteworthy Observations Interestingly, we find a cognitive interference effect. The results demonstrate that cognitive guidance significantly enhances human detection performance for high-quality videos, such as those generated by higher-performance Veo3.1-Fast model. Take Figure 1 (b) as an example, AI-generated videos are already extremely similar to real videos in terms of low-level image features and other factors, human direct intuitive judgment often fail to distinguish them. Cognitive guidance in these cases encourages humans to analyze the plausibility of the video, such as object shapes and adherence to physical laws. However, a different pattern emerges for low-quality videos. Specifically, the introduction of lower-level cognitive guidance may cause humans to overly doubt the authenticity of originally real video content, which temporarily leads to a decrease in *SR*. Take

Figure 1 (a) as an example, humans may misjudge a video due to perceived visual blurriness or disappearance of hands. But higher-level cognitive guidance helps humans to use logical reasoning to analyze the causal plausibility of video content, thereby reducing the interference caused by excessive sensory focus. Thus, our first core finding is:

Finding 1: Cognitive guidance can effectively improve human detection performance, whereas for low-quality videos, lower-level guidance may interfere with detection performance.

4.2 How Models Performs in Detection

Model Setting To reproduce model-based detection performance, we select open-sourced models, including the LLaVA series (LLaVA-NeXT-Video-7B and Video-LLaVA-7B), the Qwen series (Qwen2.5-VL-3B, Qwen3-VL-2B, Qwen3-VL-4B, Qwen3-VL-8B, and Qwen3-VL-32B), the unified model (Vila-U-7b and Show-o2-1.5B), and the specialized long video understanding model Video-XL2-8B. We make the evaluation settings of MLLMs consistent with human evaluators across various cognitive scenarios, and evaluation prompts for MLLMs are provided in Appendix D.

Role of Video Understanding Capabilities As shown in Figure 2 (b), the cognitive interference effect is also observed in model-based detection. Slightly different from humans, cognitive guidance gradually shifts the model's judgment logic from analyzing single-frame image quality to relying on the model's own understanding of the continuous temporal content in videos. Furthermore, we investigated the impact of the model's video understanding capabilities on the effect of cognitive guidance. Limitations in the model's video understanding capabilities reduce detection performance. As shown in Figure 2 (c), due to its inherent deficit in semantic understanding, the Vila-U-7B model fails to effectively integrate guidance related to narrative continuity and scene interaction into its detection process in the **Semantics and Narrative Scenario**. And better model video reasoning and understanding capabilities lead to improved detection performance. Represented by Qwen3-VL-8B as shown in Figure 2 (d), Qwen series models, relying on their strong understanding of physical laws, are able to effectively leverage guidance related to the realism of motion, object stability, and other physical phenomena in the **Physical Plausibility Scenario**, leading to a significant improvement in detection performance. Thus, we can have our second core finding shown as follows:

Finding 2: For models with stronger understanding and logical reasoning capabilities, their detection performance improves as the cognitive information in the scenario progresses from low-level visual perception to high-level logical reasoning.

4.3 Humans vs. MLLMs Analysis

Comparing the detection performance of humans and models, we find that humans generally perform better than models when relying on intuitive judgment. However, this gap becomes smaller after cognitive guidance is introduced. Taking the subset containing videos generated by the Hunyuan model as an example, In the **Zero-Knowledge Scenario**, humans achieve 75.00% *SR* compared to 50.00% for Vila-U-7B. In the **Social and Cultural Scenario**, humans achieve 81.20% versus 75.00% for Vila-U-7B. This indicates models may lack a clear direction for analysis. When structured guidance provides this direction, models can quickly improve their detection performance. As a result, the performance gap between humans and models may continue to shrink and would eventually disappear. This leads to our third core finding:

Finding 3: The detection performance gap between humans and models narrows under cognitive information-based detection guidance.

4.4 Factors Behind Model-Based Detection Performance

Using samples from the CogDetect dataset, we mainly analyze three factors: model size, model paradigm, and Scenario complexity. Related prompts can be found in Appendix D. The experimental results are shown in Figure 3, and more detailed data can be found in Appendix E.

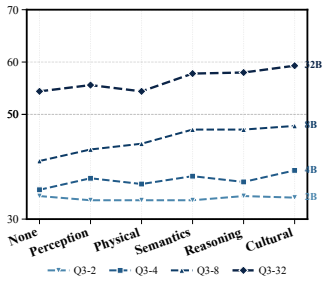

Figure (a) Model Size

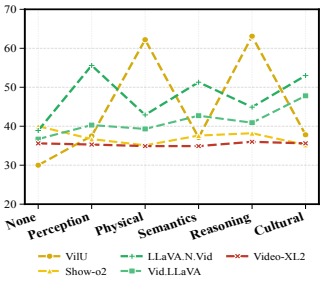

Figure (b) Model Paradigm

Table (c) Scenario complexity. This table shows the *SR* under three scenarios that we designed: **Simple Scenario**, **Basic Scenario** and **Complex Scenario**.

Scenario	Simple	Basic	Complex
None	50.71	57.88	56.79
Perception	49.29	55.54	55.09
Physical	48.93	55.57	52.14
Semantics	46.79	55.87	54.36
Reasoning	47.50	57.77	55.00
Culture	49.64	57.38	54.05

Figure 3: **Factors Behind Model-Based Detection Performance.** Figure (a) and Figure (b) shows the detection performance of different model sizes and paradigms across various scenarios. Table (c) presents model-based detection performance across different scenario complexities.

Factor 1: Model Size Figure 3 (a) shows detection performance of different model sizes in the Qwen3-VL series. We observe the larger the size, the more effectively cognitive guidance can be utilized. This is because large-scale model, such as the 32B model, has strong general understanding and complex logical reasoning capabilitiesYang et al. (2025).

Factor 2: Model Paradigm As shown in Figure 3 (b), we find that the impact of cognitive guidance varies with different model paradigms. Compared to the stable detection performance of the Qwen series models across various scenarios, the LLaVA series models are more sensitive to different cognitive information guidance, resulting in larger variations in detection performance. Detection performance of unified models remains with no clear correlation pattern emerging across various cognitive guidance. Notably, the Video-XL2 model, specifically designed for long video understanding, shows almost no change in *SR*. This is likely due to its VST-based architecture Shu et al. (2025), which provides robust spatiotemporal understanding and makes it largely unaffected by external instructions.

Factor 3: Scenario complexity The sensitivity to prompts is the most significant factor influencing the final model-based detection performance. Concretely, we design simple and complex scenarios based on existing ones. As shown in Table 3 (c), in the **Simple Scenario** retaining only the basic elements needed for judgment with brief guidance(e.g., "Carefully analyze from the Perception and Senses aspect."), the model lacks sufficient judgment basis, and the effect of guidance on detection performance is not significant. As the complexity increases, in the **Basic Scenario**, more detailed descriptions of cognitive information are introduced(e.g., adding specific descriptions of the Perception and Senses Scenario and Visual Quality within the scenario), making the semantic information richer and enabling the model to perform better reasoning and judgment. However, when the scenario becomes too complex(e.g., with additional information such as roles, styles, and more detailed cognitive information description), it may introduce distractions and cause the model's attention to be diverted, leading to a decline in detection performance.

5 Iterative Cognitive-Layered Instruction Evolution

Motivated by our observation that different cognitive scenarios require distinct guidance strategies, and that models exhibit high sensitivity to detection instructions, we propose an instance-driven, cognitive-layered instruction self-evolution framework. Rather than relying on manually engineered prompts, the framework enables a multimodal model to iteratively refine scenario-specific detection instructions through structured self-reflection grounded in our cognitive hierarchy.

5.1 Scenario-Aware Instruction Evolution Framework

Let $\mathcal{M}$ denote a pretrained multimodal model and $\{v_1, \ldots, v_N\}$ a small set of AI-generated videos with known ground truth. For each cognitive scenario $s \in \mathcal{S}$, we maintain a scenario-specific instruction I_s initialized from human cognitive priors defined in our CogDetect framework (Sec. 3). At each iteration k, rather than performing binary classification, $\mathcal{M}$ is prompted to analyze v_k and assess the adequacy of I_s itself under the lens of scenario s. Accordingly, Φ denotes the multimodal model used to extract informative content beneficial for video detection across different scenarios. The instruction is updated via an evolution operator $\mathcal{E}_s$:

$$I_s^{(k+1)} = \mathcal{E}_s\big(I_s^{(k)}, v_k\,;\,\mathcal{M}\big), \tag{1}$$

Input: Pretrained multimodal model $\mathcal{M}$; AI-generated video set $\mathcal{V} = \{v_1, \ldots, v_N\}$; cognitive scenarios $\mathcal{S}$; initial instructions $\{I_s^{(0)}\}$.

Output: Evolved detection instructions $\{I_s^*\}$.

for $s \in \mathcal{S}$ **do**
  $I_s \leftarrow I_s^{(0)}$ ▷ Initialize scenario instruction
  for $k = 1, \ldots, N$ **do**
    $A_{s,k} \leftarrow \mathcal{M}(v_k, I_s)$ ▷ Model-guided video analysis
    $\Delta_{s,k} \leftarrow \Phi(A_{s,k})$ ▷ Extract AI-specific artifacts
    $I_s \leftarrow \mathcal{E}_s(I_s, \Delta_{s,k})$ ▷ Instruction refinement
  end for
  $I_s^* \leftarrow I_s$ ▷ Store evolved instruction
end for

Algorithm 1: Scenario-Aware Instruction Self-Evolution for AI-Generated Video Detection

where each application of $\mathcal{E}_s$ involves three constrained steps anchored to the cognitive scenario: (i) identifying artifacts in v_k relevant to scenario s (e.g., physical implausibilities, semantic inconsistencies); (ii) diagnosing which criteria in $I_s^{(k)}$ fail to capture these artifacts; and (iii) refining the instruction by sharpening existing criteria without rewriting it from scratch, preserving cognitive alignment.

Iterating over diverse AI-generated videos drives the instruction from instance-specific cues toward generalizable detection principles within each cognitive layer. After N iterations, the evolved instruction is:

$$I_s^* = \mathcal{E}_s^{(N)} \circ \cdots \circ \mathcal{E}_s^{(1)} \left(I_s^{(0)} \right), \tag{2}$$

where $I_s^{(0)}$ is the cognitive prior-based initialization and I_s^* is the converged, model-aligned instruction for scenario s.

Evolution is conducted independently per cognitive scenario to allow each instruction to specialize within its designated reasoning level. The resulting instructions are then integrated into a unified detection prompt that preserves the hierarchical structure of our cognitive framework while eliminating redundancy across scenarios.

5.2 Experiments: Effectiveness and Generalizability of The Framework

We analyze the proposed framework from two perspectives: effectiveness and generalizability. We use Qwen2.5-3B as the base model for evolution and select 7 videos of varying quality levels as scenario examples. The model iteratively refines instructions across five cognitive scenarios, yielding a unified evolved instruction that integrates multi-level cognitive guidance. All evaluations are conducted on the GenVideo-val dataset Zhang et al. (2025a), which is a standard benchmark for AI-generated video detection. Additional experimental details and results are provided in Appendix F.

Effectiveness. As shown in Figure 4, the evolved instruction substantially improves detection performance on the base model, consistently achieving higher accuracy than the initial instruction. This confirms that instruction self-evolution enables the model to internalize more discriminative, model-aligned detection cues, shifting from instance-level observations to stable and generalizable detection principles. Furthermore, **Base Prompt** serves as a baseline consisting of initial detection instructions and basic video context without the guidance of our cognitively layered evolution. Results demonstrate that evolved instruction significantly enhances the model's judgment accuracy in complex scenarios.

We further compare against recent training-based detectors, including DeMamba Chen et al. (2024b), NPR Tan et al. (2024), TALL Xu et al. (2023), STIL Gu et al. (2021), and NSG-VD Zhang et al. (2025a), using the same evaluation settings as NSG-VD Zhang et al. (2025a) (details in Appendix F). As shown in Table 1, despite being entirely training-free, our method achieves 77.09% *SR* and 94.15% *AUC*, competitive with fully supervised detectors, demonstrating that instruction-level evolution can effectively narrow the gap with conventional supervised systems.

Table 1: Comparison with training-based video detection methods. Competing models require large-scale training. In contrast, our instruction evolution framework is entirely training-free. Despite requiring no parameter optimization, our method achieves competitive performance, demonstrating the effectiveness of instruction-level optimization.

Method	DeMamba	NPR	TALL	STIL	NSG-VD	Base Prompt	Evolution (Ours)
SR	84.21	77.96	79.85	63.51	91.46	52.7	77.09
Recall	72.02	57.35	60.87	27.02	88.02	48.1	74.41
AUC	93.88	93.02	95.67	93.49	96.14	84.6	94.15

Generalizability. *Cross-model generalization.* We apply the evolved instruction directly to multiple MLLMs on the cogDetect dataset without further adaptation. As shown in Figure 4, it consistently improves detection across different architectures, indicating that the evolved instruction captures broadly transferable detection cues rather than overfitting to the base model. We attribute this to the cognitively layered evolution mechanism, which integrates complementary signals from low-level perception to high-level reasoning, aligning well with shared multimodal representations across models.

Cross-data generalization. Although evolution is conducted on videos from specific generators, the evolved instruction maintains strong performance on unseen generation sources with distinct artifact characteristics. This suggests the instruction encodes fundamental inconsistencies common across AI-generated videos, such as disruptions in temporal coherence, physical plausibility, and semantic consistency, rather than generator-specific cues, enabling robust generalization to previously unseen generative mechanisms.

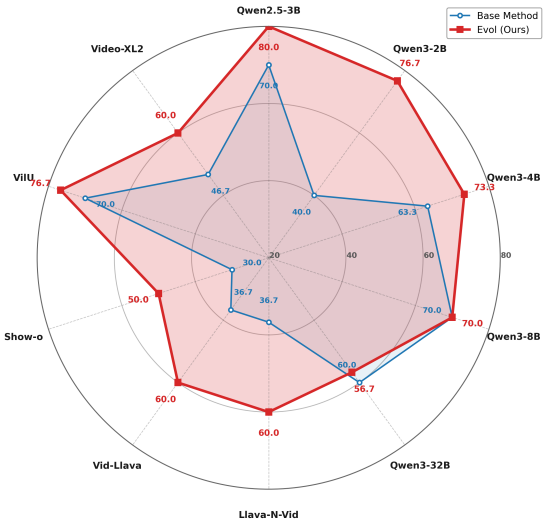

Figure 4: Comparison between the proposed instruction evolution framework and the base instruction. All methods are evaluated on the same video dataset, including both real and AI-generated videos. We report detection *SR* across multiple multimodal models. Our evolved instruction outperforms the base instruction and demonstrates strong generalization.

6 Conclusion

This paper presents a cognitive-layered framework for AI-generated video detection, investigating how progressively higher-level cognitive guidance affects multimodal models beyond low-level visual artifacts. We first establish a hierarchical framework of cognitive scenarios based on fundamental human cognitive processes, then conduct experimental analyses to determine how cognitive guidance impacts the detection performance of both humans and MLLMs. Based on our findings, we propose an iterative, instance-driven, cognitive-layered instruction self-evolution framework. This training-free evolution method incorporates cognitive guidance into the detection process, improving the ability of MLLMs to interpret video content and detect videos without requiring additional fine-tuning. Furthermore, our experiments demonstrate that this approach generalizes effectively to different models and diverse sets of unseen video data.

Limitations Based on our proposed evolutionary detection framework, MLLMs can leverage cognitive information guidance to detect AI-generated videos scalably and efficiently, without requiring large-scale annotated training data or model fine-tuning. However, the effectiveness of this framework remains fundamentally bounded by the video understanding capabilities of MLLMs. In the future, we may further explore how human cognitive processes can be more deeply integrated into the pretraining or instruction-tuning stage of MLLMs, with the goal of enhancing their foundational video understanding capabilities rather than compensating for their deficiencies at inference time.

AI USE STATEMENT

In the preparation of this work, large language models (LLMs) were used for the following purposes: (1) polishing and refining the language of the manuscript; (2) inspecting and debugging experimental code; and (3) cross-checking experimental results and numerical consistency. LLMs were **not** used for formulating research ideas, designing experiments or methodologies, generating or fabricating experimental data, or drafting scientific content. All AI-assisted outputs were manually reviewed, verified, and (where applicable) corrected by the authors. The authors take full responsibility for the integrity and content of this work.

REFERENCES

George A Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends in cognitive sciences*, 15(3):122–131, 2011.

Jianfa Bai, Man Lin, and Gang Cao. Ai-generated video detection via spatio-temporal anomaly learning.(2024). *arXiv preprint arXiv:2403.16638*, 2024.

Xu Cao, Yifan Shen, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Meihuan Huang, Jianguo Cao, Aidong Zhang, et al. What is the visual cognition gap between humans and multimodal llms? *arXiv preprint arXiv:2406.10424*, 2024.

Patrick Cavanagh. Visual cognition. *Vision research*, 51(13):1538–1551, 2011.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024a.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Lin Geng Foo, Hossein Rahmani, and Jun Liu. Ai-generated content (aigc) for various data modalities: A survey. *ACM Computing Surveys*, 57(9):1–66, 2025.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3473–3481, 2021.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

Peisong He, Leyao Zhu, Jiaxing Li, Shiqi Wang, and Haoliang Li. Exposing ai-generated videos: A benchmark dataset and a local-and-global temporal defect based detection method. *arXiv preprint arXiv:2405.04133*, 2024.

Glyn W Humphreys and Vicki Bruce. *Visual cognition: Computational, experimental and neuropsychological perspectives*. Psychology Press, 2021.

Haojian Jin, Yale Song, and Koji Yatani. Elasticplay: Interactive video summarization with dynamic time budgets. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1164–1172, 2017.

Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, and Feng Xia. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57(6):159, 2024.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020.

Chang Liu and Han Yu. Ai-empowered persuasive video generation: A survey. *ACM Computing Surveys*, 55(13s):1–31, 2023.

Xinxin Liu, Zhaopan Xu, Ming Li, Kai Wang, Yong Jae Lee, and Yuzhang Shang. Can world simulators reason? gen-vire: A generative visual reasoning benchmark. *arXiv preprint arXiv:2511.13853*, 2025a.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024.

Yuanxin Liu, Rui Zhu, Shuhuai Ren, Jiacong Wang, Haoyuan Guo, Xu Sun, and Lu Jiang. Uve: Are mllms unified evaluators for ai-generated videos? *arXiv preprint arXiv:2503.09949*, 2025b.

Long Ma, Jiajia Zhang, Hongping Deng, Ningyu Zhang, Yong Liao, and Haiyang Yu. Decof: Generated video detection via frame consistency. *CoRR*, 2024.

OpenAI. Video generation models as world simulators, 2024.

Razvan Petrusel, Jan Mendling, and Hajo A Reijers. How visual cognition influences process model comprehension. *Decision Support Systems*, 96:1–16, 2017.

Markus Schatten. Ai and the future of entertainment technology. 2024.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26160–26169, 2025.

Christopher Summerfield and Tobias Egner. Expectation (and attention) in visual cognition. *Trends in cognitive sciences*, 13(9):403–409, 2009.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. *Advances in Neural Information Processing Systems*, 37:52723–52748, 2024.

Shah Muhammad Imtiyaj Uddin, Rashedul Islam Sumon, Md Ariful Islam Mozumder, Md Kamran Hussin Chowdhury, Tagne Poupi Theodore Armand, and Hee Cheol Kim. Innovations and challenges of ai in film: a methodological framework for future exploration. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(7):1–55, 2025.

Dingrui Wang, Hongyuan Ye, Zhihao Liang, Zhexiao Sun, Zhaowei Lu, Yuchen Zhang, Yuyu Zhao, Yuan Gao, Marvin Seegert, Finn Schäfer, et al. Target-bench: Can world models achieve mapless path planning with semantic targets? *arXiv preprint arXiv:2511.17792*, 2025a.

Jiaqi Wang, Weijia Wu, Yi Zhan, Rui Zhao, Ming Hu, James Cheng, Wei Liu, Philip Torr, and Kevin Qinghong Lin. Video reality test: Can ai-generated asmr videos fool vlms and humans? *arXiv preprint arXiv:2512.13281*, 2025b.

Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18869–18880, 2025c.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Shuai Wang, Daoan Zhang, Tianyi Bai, Shitong Shao, Jiebo Luo, and Jiaheng Wei. Last: Learning to think in space and time for generalist vision-language models. *arXiv preprint arXiv:2511.19261*, 2025d.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025e.

Zeqing Wang, Xinyu Wei, Bairui Li, Zhen Guo, Jinrui Zhang, Hongyang Wei, Keze Wang, and Lei Zhang. Videoverse: How far is your t2v generator from a world model? *arXiv preprint arXiv:2510.08398*, 2025f.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22658–22668, 2023.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18826–18836, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for ai-generated video detection. *arXiv preprint arXiv:2510.08073*, 2025a.

Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. Q-bench-video: Benchmark the video quality understanding of lmms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3229–3239, 2025b.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

APPENDIX

A VISUALIZATION

In this section, which corresponds to Section 3 of the main text, we provide a comprehensive visualization of the five cognitive scenarios and different types of videos.

A.1 VISUALIZATION OF FIVE COGNITIVE SCENARIOS

We provide a detailed visualization of video frames that reflect various types of cognitive information. These include perception and senses, physical plausibility, semantics and narrative, reasoning and logic, and social and cultural scenarios. By presenting these representative video sequences, we intuitively demonstrate the visual characteristic issues across different dimensions in various videos. The detailed visualizations are presented in Figure 5.

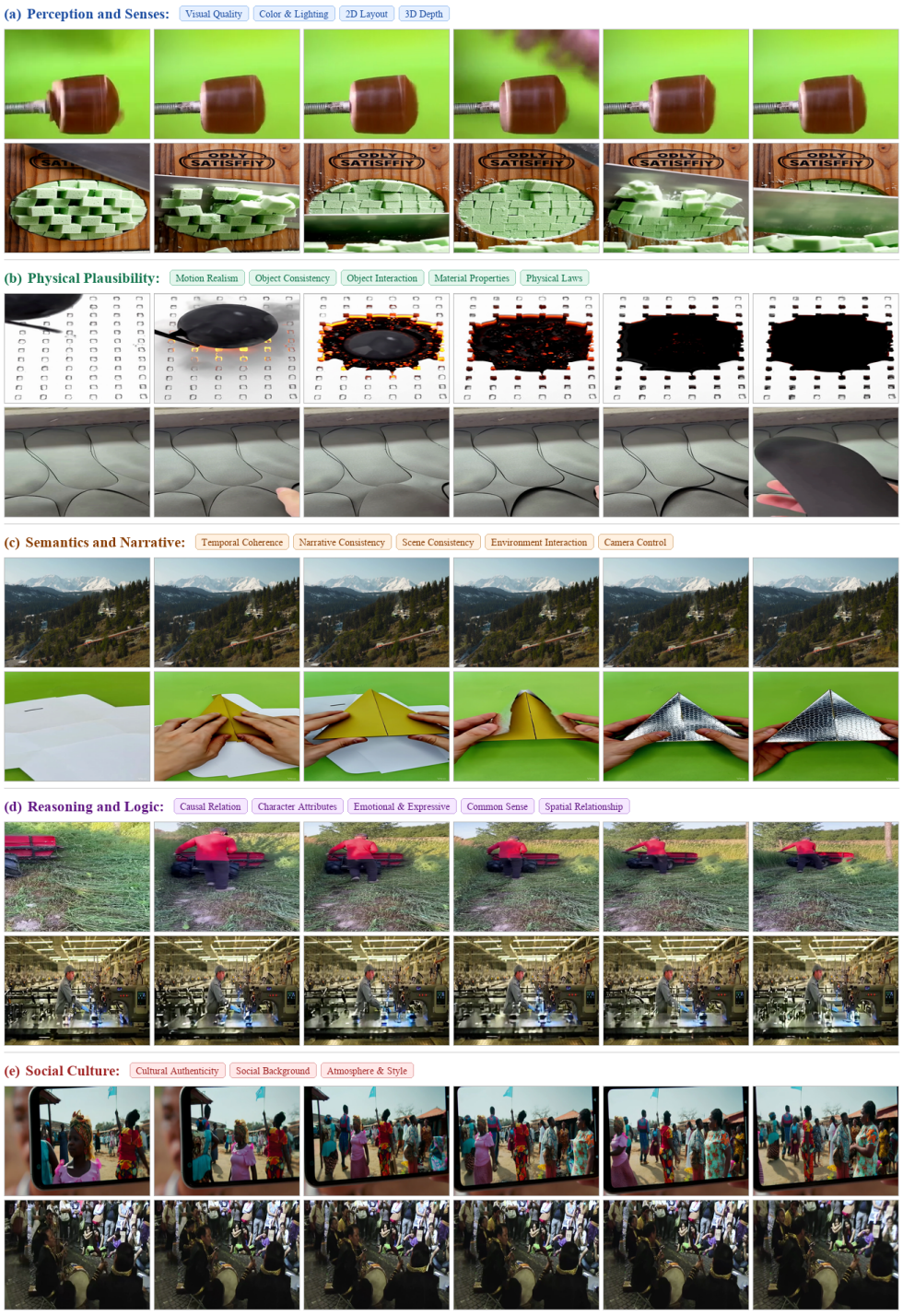

Figure 5: Extended visualization of Five Cognitive Scenarios. This figure provides frame sequences for each of the five cognitive scenarios defined in our framework (**Perception and Senses Scenarios**, **Physical Plausibility Scenarios**, **Semantics and Narrative Scenarios**, **Reasoning and Logic Scenarios**, and **Social and Cultural Scenarios**).

A.2 VISUALIZATION OF DIFFERENT TYPES OF VIDEOS

In this section, we further visualize video frames of varying quality generated by different video generation models. These models include *StepVideo*, *OpenSora-V2*, *Sora2*, *HunyuanVideo*, *Wan2.2*, and *Veo3.1-Fast*. Based on their visual effects, we categorize the videos produced by these models into three types: low, medium, and high quality. The detailed visualizations are presented in Figure 6.

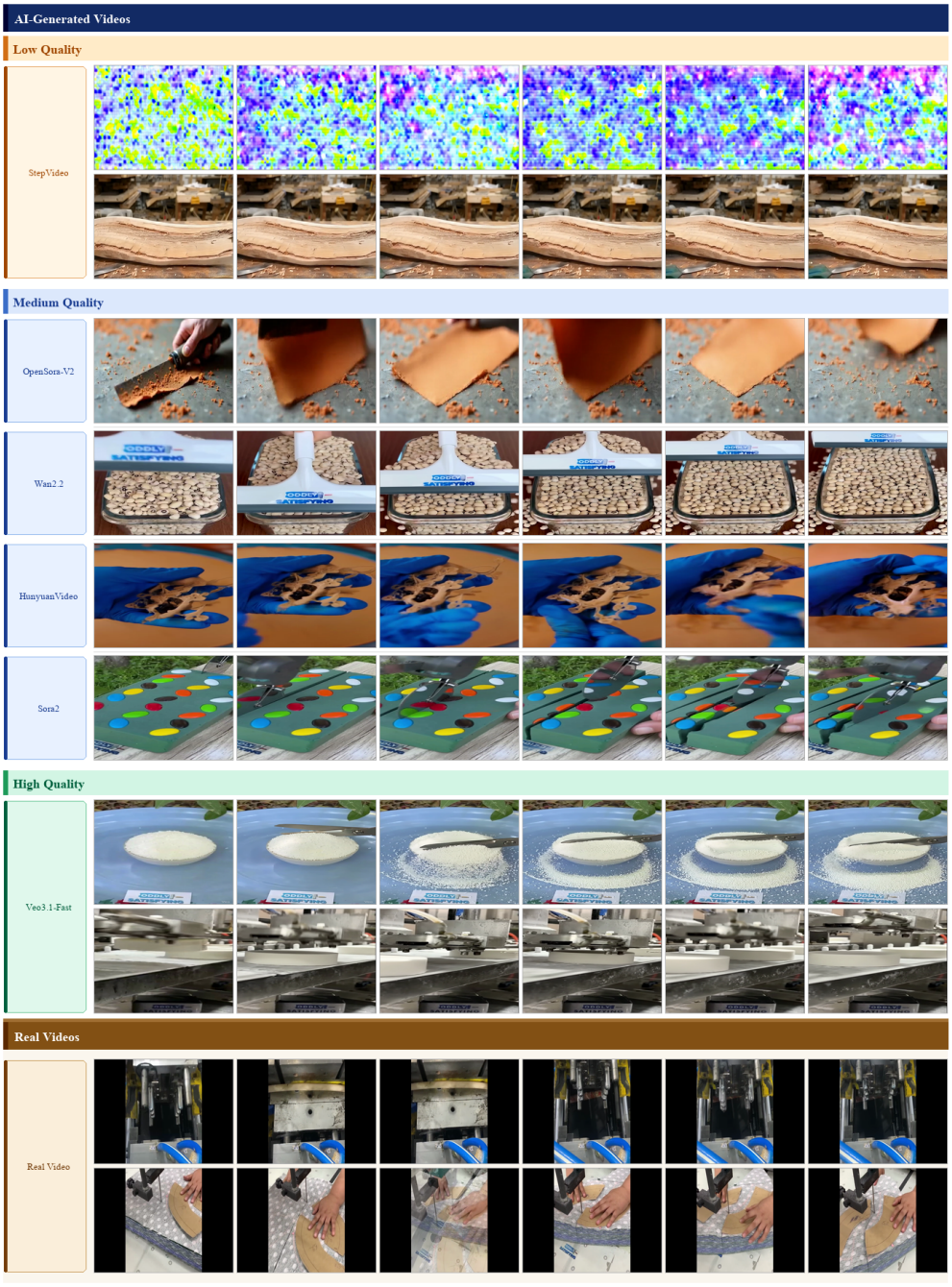

Figure 6: visualization of different types of videos. This figure displays frame sequences from different types of videos, encompassing real video and AI-generated samples across three quality tiers (low, medium, and high) derived from six distinct generative models.

B Details and Results of Human vs. MLLM Detection

In this section, which corresponds to Section 4 of the main text, we extend the analysis presented in the primary study by providing a more detailed evaluation of detection performance. Appendix B.1 describes the experimental setup. Building on this, Appendix B.2 shows the SR. In Appendices B.3 and B.4, we further examine the False Negative Rate (FNR) and False Positive Rate (FPR) metrics to compare the judgment logic of humans and multimodal large language models in AI-generated video detection tasks. In Appendices B.5 and B.6, we additionally report the Recall and F1-score metrics. Additionally, Appendix B.7 presents the specific numerical results of detection performance across six different video generation models. By visualizing and analyzing these metrics, we aim to identify the particular cognitive scenarios and generative sources where current detection methods encounter significant challenges.

B.1 Experiment Setup

Details about the constructed dataset. As shown in Table 2, we constructed the CogDetect dataset by selecting 2,411 videos from VideoRealityTest, UVE Bench, Kinetics 400, and DeeptraceReward. In addition, we used the GenVideo-val dataset to evaluate the effectiveness and generalizability of our framework, which is a standard benchmark for AI-generated video detection. As shown in Table 3, the human evaluation used 341 videos, including 264 AI-generated videos and 77 real videos. Specifically, We randomly selected 44 generated videos from each generative model: *StepVideo, OpenSora V2, Sora 2, HunyuanVideo, Wan 2.2, and Veo 3.1-Fast*. To match the durations of the generated videos, we selected real videos of different lengths. These included 11 videos of 4 seconds, 22 videos of 5 seconds, 22 videos of 8 seconds, and 22 videos of 10 seconds. Each subset contains 22 real videos and 44 AI-generated videos, and the ratio of real to AI-generated videos is fixed at 1:2 in each scenario.

Table 2: Dataset construction details.

Dataset / Source	**# Real**	**# AI-generated**	**Overall**
CogDetect dataset			
VideoRealityTest	54	241	295
UVE-Bench	–	283	283
Kinetics-400	243	–	243
DeeptraceReward	470	1120	1590
Overall	**767**	**1,644**	**2,411**
A standard benchmark			
GenVideo-val	1000	956	1956

Table 3: Composition of the human evaluation dataset.

AI-generated videos			**Real videos**		
Source	#	Duration	Source	#	Duration
StepVideo	44	5s	CogDetect Dataset	11	4s
OpenSora-V2	44	5s		22	5s
Sora2	44	10s		22	8s
HunyuanVideo	44	5s + 4s		22	10s
Wan2.2	44	5s		–	–
Veo3.1-Fast	44	8s		–	–
Overall	**264**	–	**Overall**	**77**	–

Evaluation Protocol. For evaluation, we treat AI-generated videos as negative samples and real videos as positive samples. Under this convention, SR measures the overall classification accuracy, FNR denotes the proportion of real videos that are incorrectly classified as AI-generated, and FPR denotes the proportion of AI-generated videos that are incorrectly classified as real. Recall measures the proportion of real videos correctly classified as real, while F1 is the harmonic mean of Precision and Recall.

Let TP denote real videos correctly classified as real, TN denote AI-generated videos correctly classified as generated, FP denote AI-generated videos classified as real, and FN denote real videos classified as generated. We compute

$$\mathrm{SR} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \quad \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}.$$

All metrics are computed before rounding, and the same label mapping is used for humans, MLLMs, and prior detectors.

B.2 SR comparison between Human and MLLM detection

Figure 7 compares the SR results of human and MLLMs under the guidance of cognitive information-based detection, with detailed analysis provided in the main text.

B.3 FNR comparison between Human and MLLM detection

The False Negative Rate (FNR) serves as a critical metric to measure the missed detection rate in AI-generated video identification tasks. As illustrated in Figure 8, we provide a detailed visualization of the FNR trends across six different video generation models with varying quality levels, comparing the performance of human observers and various detection models. The results demonstrate that as cognitive guidance progresses from low-level visual perception to high-level logical reasoning, detectors can more effectively capture deep generative traces such as violations of physical laws and inconsistent causal logic, thereby significantly reducing FNR and improving detection performance. This indicates that integrating high-level cognitive information allows models to move beyond superficial pixel-level artifacts and identify more fundamental anomalies that are often overlooked by traditional detection methods.

B.4 FPR comparison between Human and MLLM detection

The False Positive Rate (FPR) is another essential metric that reflects the probability of misidentifying real videos as AI-generated content. As illustrated in Figure 9, we analyze the FPR trends across different video sources and cognitive scenarios for both human participants and multimodal models. The results indicate that when providing excessively low-level cognitive guidance for the detection of low-quality videos, both human observers and models tend to become overly sensitive to minor imperfections. This often leads to excessive suspicion of natural details in authentic videos, such as motion blur, compression artifacts, or subtle lighting inconsistencies, thereby resulting in higher misjudgment rates and an increased FPR. This phenomenon suggests that while granular visual inspection is necessary, over-reliance on low-level cues without the balance of high-level logical context can compromise the overall reliability of the detection system by increasing false alarms.

B.5 Recall comparison between Human and MLLM detection

Recall measures the proportion of real videos correctly identified as real. As illustrated in Figure 10, we analyze the Recall trends across different video sources and cognitive scenarios for both human participants and multimodal models. The results show that Recall for both human participants and multimodal models rises overall as cognitive guidance progresses from low-level perception to high-level reasoning, confirming that higher-level guidance progressively reduces the chance of mistaking real videos for generated ones. Recall first declines and then rises again across this progression, and this fluctuation is far more pronounced in models with weaker video understanding capacity. This phenomenon suggests that while higher-level cognitive guidance benefits detection overall, only models with stronger understanding capabilities can consistently exploit it, whereas weaker models struggle to convert such guidance into a genuine detection advantage.

B.6 F1 COMPARISON BETWEEN HUMAN AND MLLM DETECTION

F1 jointly evaluates Precision and Recall for the Real class, reflecting whether a detector recognizes real videos without generating excessive false alarms. As illustrated in Figure 11, we analyze the F1 trends across different video sources and cognitive scenarios for both human participants and multimodal models. The results show that F1 for both human participants and multimodal models rises overall as cognitive guidance progresses from low-level perception to high-level reasoning, indicating that higher-level guidance helps detectors identify real videos more reliably without sacrificing precision. Human F1 improves steadily across this progression, while the models' average F1 shows the same upward trend, with models possessing stronger understanding and reasoning capabilities demonstrating more significant improvements. This phenomenon confirms our finding that for models with stronger understanding and logical reasoning capabilities, detection performance improves as the cognitive information in guidance progresses from low-level visual perception to high-level logical reasoning.

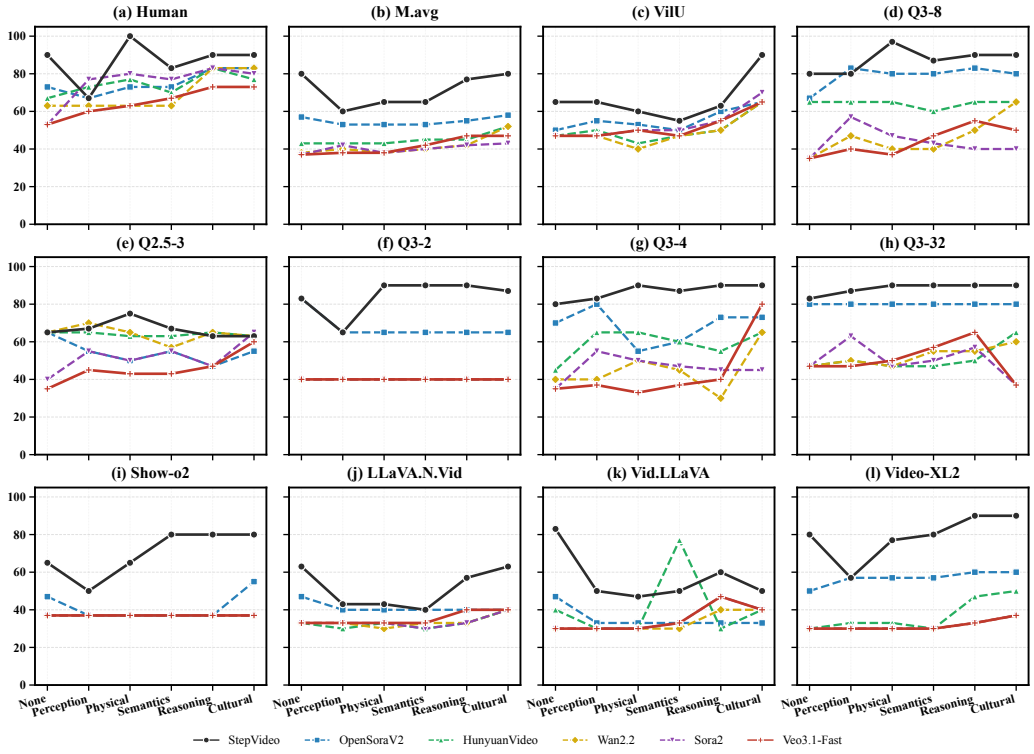

Figure 7: Comparison of Success Rate (SR) between human and MLLM detection across various scenarios.

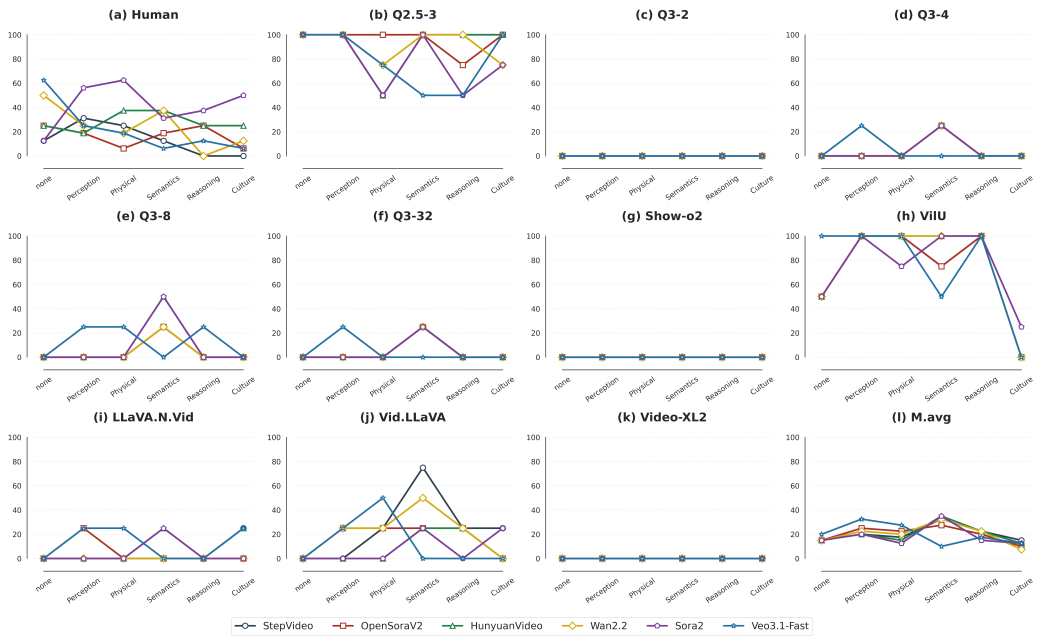

Figure 8: FNR comparison between Human and MLLM detection across various scenarios.

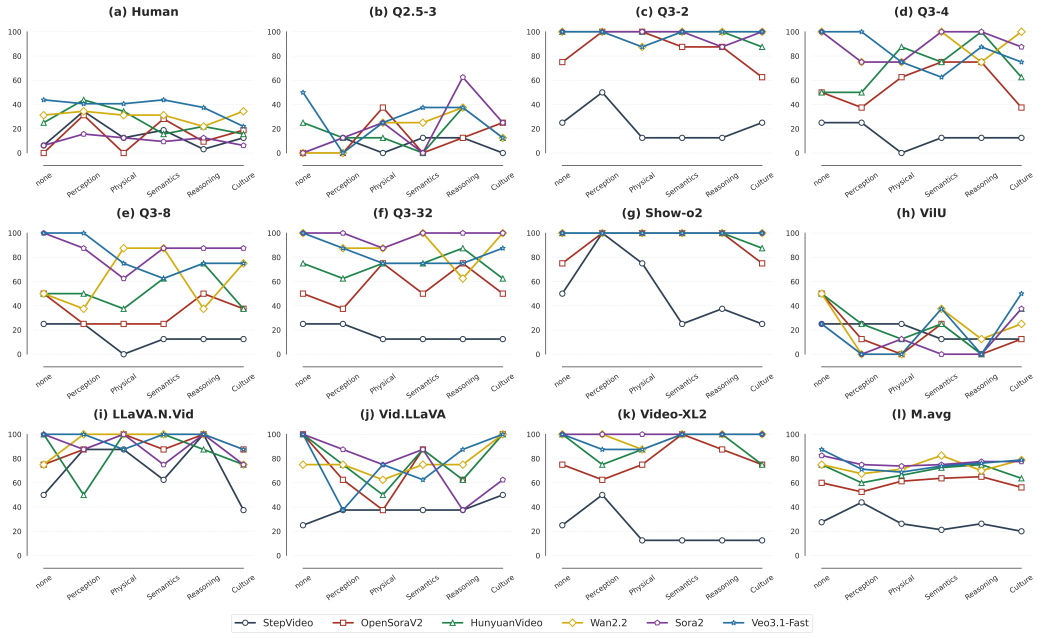

Figure 9: FPR comparison between Human and MLLM detection across various scenarios.

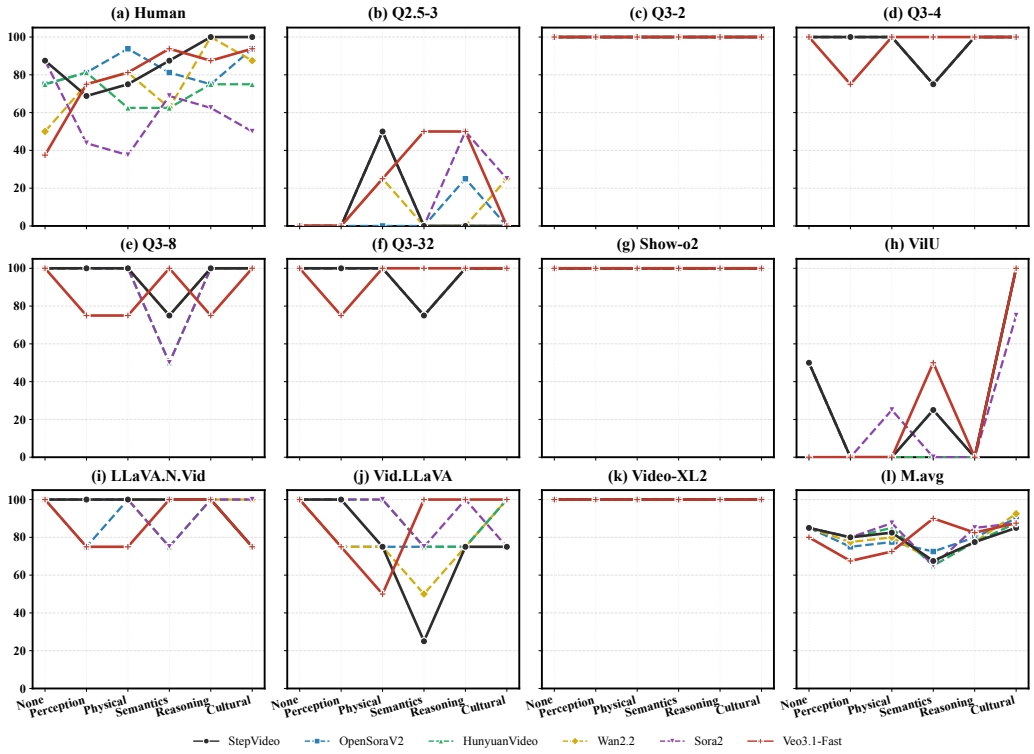

Figure 10: Recall comparison between Human and MLLM detection across various scenarios.

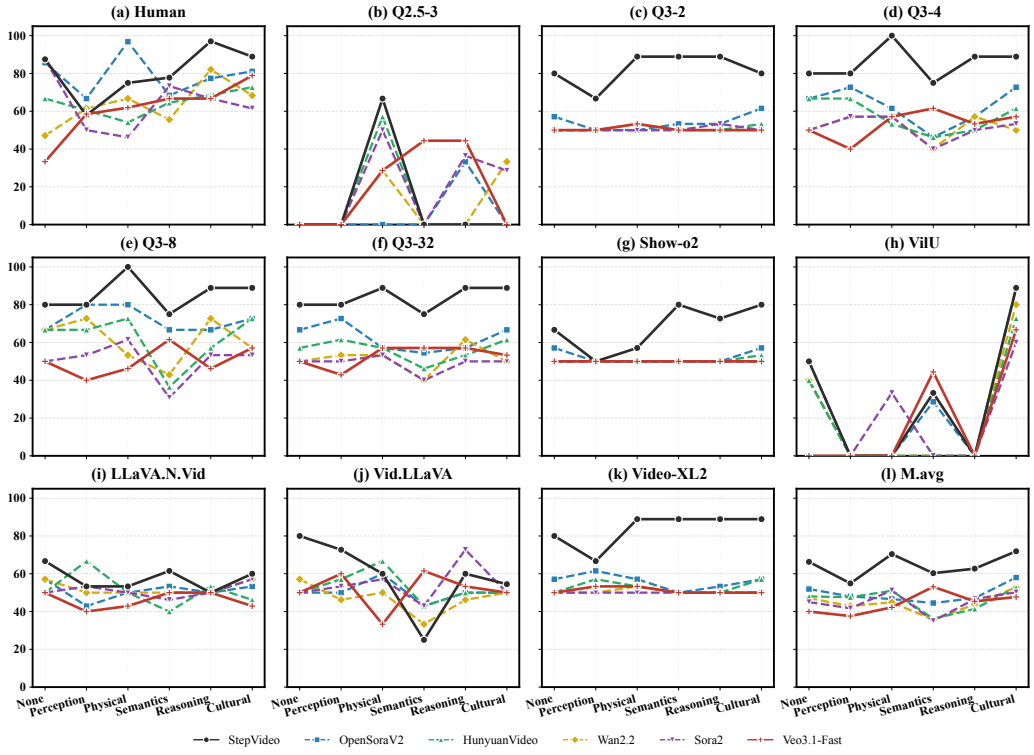

Figure 11: F1 comparison between Human and MLLM detection across various scenarios.

B.7 DETECTION RESULTS ON SUBSETS OF DIFFERENT GENERATIVE MODELS.

In this section, we present the detailed numerical results of detection performance across six different video generation models(StepVideo, HunyuanVideo, OpenSoraV2, Sora2, Wan2.2, and Veo3.1-Fast) with varying quality levels. These results are reported in Table 4, Table 5, Table 6, Table 7, Table 8, and Table 9. These tables report the Success Rate (SR), False Negative Rate (FNR), False Positive Rate (FPR), Recall (R), and F1-score (F1) for humans and MLLMs across our defined cognitive scenarios. The data includes the detection performance of human observers as well as various multimodal models across the defined cognitive scenarios. In the final row of each table, the term M.avg represents the average numerical results calculated across all different models. This average value is specifically used to provide a comprehensive comparison between the overall performance of the models and the detection results of human participants.

Table 4: Detailed detection performance on the StepVideo subset, categorized by cognitive scenarios.

Model	Indicator	None	Perception	Physical	Semantics	Reasoning	Culture
Human	SR	91.7	66.7	83.3	83.3	97.9	91.7
	FNR	12.5	31.2	25.0	12.5	0.0	0.0
	FPR	6.2	34.4	12.5	18.8	3.1	12.5
	R	87.5	68.8	75.0	87.5	100.0	100.0
	F1	87.5	57.9	75.0	77.8	97.0	88.9
Q2.5-3	SR	66.7	58.3	83.3	58.3	58.3	66.7
	FNR	100.0	100.0	50.0	100.0	100.0	100.0
	FPR	0.0	12.5	0.0	12.5	12.5	0.0
	R	0.0	0.0	50.0	0.0	0.0	0.0
	F1	0.0	0.0	66.7	0.0	0.0	0.0
Q3-2	SR	83.3	66.7	91.7	91.7	91.7	83.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	25.0	50.0	12.5	12.5	12.5	25.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	80.0	66.7	88.9	88.9	88.9	80.0
Q3-4	SR	83.3	83.3	100.0	83.3	91.7	91.7
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	25.0	25.0	0.0	12.5	12.5	12.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	80.0	80.0	100.0	75.0	88.9	88.9
Q3-8	SR	83.3	83.3	100.0	83.3	91.7	91.7
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	25.0	25.0	0.0	12.5	12.5	12.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	80.0	80.0	100.0	75.0	88.9	88.9
Q3-32	SR	83.3	83.3	91.7	83.3	91.7	91.7
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	25.0	25.0	12.5	12.5	12.5	12.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	80.0	80.0	88.9	75.0	88.9	88.9
Show-o2	SR	66.7	33.3	50.0	83.3	75.0	83.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	50.0	100.0	75.0	25.0	37.5	25.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	66.7	50.0	57.1	80.0	72.7	80.0
VilU	SR	66.7	50.0	50.0	66.7	58.3	91.7
	FNR	50.0	100.0	100.0	75.0	100.0	0.0
	FPR	25.0	25.0	25.0	12.5	12.5	12.5
	R	50.0	0.0	0.0	25.0	0.0	100.0
	F1	50.0	0.0	0.0	33.3	0.0	88.9
LLaVA.N.Vid	SR	66.7	41.7	41.7	58.3	33.3	66.7
	FNR	0.0	0.0	0.0	0.0	0.0	25.0
	FPR	50.0	87.5	87.5	62.5	100.0	37.5
	R	100.0	100.0	100.0	100.0	100.0	75.0
	F1	66.7	53.3	53.3	61.5	50.0	60.0
Vid.LLaVA	SR	83.3	75.0	66.7	50.0	58.3	50.0
	FNR	0.0	0.0	25.0	75.0	25.0	25.0
	FPR	25.0	37.5	37.5	37.5	37.5	50.0
	R	100.0	100.0	75.0	25.0	75.0	75.0
	F1	80.0	72.7	60.0	25.0	60.0	54.5
Video-XL2	SR	83.3	66.7	91.7	91.7	91.7	91.7
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	25.0	50.0	12.5	12.5	12.5	12.5
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	80.0	66.7	88.9	88.9	88.9	88.9
M.avg	SR	76.7	64.2	76.7	75.0	74.2	80.8
	FNR	15.0	20.0	17.5	32.5	22.5	15.0
	FPR	27.5	43.8	26.2	21.2	26.2	20.0
	R	85.0	80.0	82.5	67.5	77.5	85.0
	F1	66.3	54.9	70.4	60.3	62.7	71.9

Table 5: Detailed detection performance on the HunyuanVideo subset, categorized by cognitive scenarios.

Model	Indicator	None	Perception	Physical	Semantics	Reasoning	Culture
Human	SR	75.0	64.6	64.6	77.1	77.1	81.2
	FNR	25.0	18.8	37.5	37.5	25.0	25.0
	FPR	25.0	43.8	34.4	15.6	21.9	15.6
	R	75.0	81.2	62.5	62.5	75.0	75.0
	F1	66.7	60.5	54.1	64.5	68.6	72.7
Q2.5-3	SR	50.0	58.3	75.0	66.7	41.7	58.3
	FNR	100.0	100.0	50.0	100.0	100.0	100.0
	FPR	25.0	12.5	12.5	0.0	37.5	12.5
	R	0.0	0.0	50.0	0.0	0.0	0.0
	F1	0.0	0.0	57.1	0.0	0.0	0.0
Q3-2	SR	33.3	33.3	33.3	33.3	33.3	41.7
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	100.0	100.0	100.0	87.5
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	50.0	50.0	50.0	53.3
Q3-4	SR	66.7	66.7	41.7	41.7	33.3	58.3
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	50.0	50.0	87.5	75.0	100.0	62.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	66.7	66.7	53.3	46.2	50.0	61.5
Q3-8	SR	66.7	66.7	75.0	41.7	50.0	75.0
	FNR	0.0	0.0	0.0	50.0	0.0	0.0
	FPR	50.0	50.0	37.5	62.5	75.0	37.5
	R	100.0	100.0	100.0	50.0	100.0	100.0
	F1	66.7	66.7	72.7	36.4	57.1	72.7
Q3-32	SR	50.0	58.3	50.0	41.7	41.7	58.3
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	75.0	62.5	75.0	75.0	87.5	62.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	57.1	61.5	57.1	46.2	53.3	61.5
Show-o2	SR	33.3	33.3	33.3	33.3	33.3	41.7
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	100.0	100.0	100.0	87.5
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	50.0	50.0	50.0	53.3
VilU	SR	50.0	50.0	58.3	50.0	66.7	75.0
	FNR	50.0	100.0	100.0	100.0	100.0	0.0
	FPR	50.0	25.0	12.5	25.0	0.0	37.5
	R	50.0	0.0	0.0	0.0	0.0	100.0
	F1	40.0	0.0	0.0	0.0	0.0	72.7
LLaVA.N.Vid	SR	33.3	66.7	33.3	25.0	41.7	41.7
	FNR	0.0	0.0	0.0	25.0	0.0	25.0
	FPR	100.0	50.0	100.0	100.0	87.5	75.0
	R	100.0	100.0	100.0	75.0	100.0	75.0
	F1	50.0	66.7	50.0	40.0	53.3	46.2
Vid.LLaVA	SR	33.3	41.7	33.3	33.3	50.0	33.3
	FNR	0.0	0.0	0.0	25.0	25.0	0.0
	FPR	100.0	75.0	50.0	87.5	62.5	100.0
	R	100.0	100.0	100.0	75.0	75.0	100.0
	F1	50.0	57.1	66.7	42.9	50.0	50.0
Video-XL2	SR	33.3	50.0	41.7	33.3	33.3	50.0
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	75.0	87.5	100.0	100.0	75.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	57.1	53.3	50.0	50.0	57.1
M.avg	SR	45.0	52.5	47.5	40.0	42.5	53.3
	FNR	15.0	20.0	15.0	35.0	22.5	12.5
	FPR	75.0	60.0	66.2	72.5	75.0	63.7
	R	85.0	80.0	85.0	65.0	77.5	87.5
	F1	48.0	47.6	51.0	36.2	41.4	52.8

Table 6: Detailed detection performance on the OpenSoraV2 subset, categorized by cognitive scenarios.

Model	Indicator	None	Perception	Physical	Semantics	Reasoning	Culture
Human	SR	91.7	72.9	97.9	75.0	85.4	85.4
	FNR	25.0	18.8	6.2	18.8	25.0	6.2
	FPR	0.0	31.2	0.0	28.1	9.4	18.8
	R	75.0	81.2	93.8	81.2	75.0	93.8
	F1	85.7	66.7	96.8	68.4	77.4	81.1
Q2.5-3	SR	66.7	66.7	41.7	66.7	66.7	50.0
	FNR	100.0	100.0	100.0	100.0	75.0	100.0
	FPR	0.0	0.0	37.5	0.0	12.5	25.0
	R	0.0	0.0	0.0	0.0	25.0	0.0
	F1	0.0	0.0	0.0	0.0	33.3	0.0
Q3-2	SR	50.0	33.3	33.3	41.7	41.7	58.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	75.0	100.0	100.0	87.5	87.5	62.5
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	57.1	50.0	50.0	53.3	53.3	61.5
Q3-4	SR	66.7	75.0	58.3	41.7	50.0	75.0
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	50.0	37.5	62.5	75.0	75.0	37.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	66.7	72.7	61.5	46.2	57.1	72.7
Q3-8	SR	66.7	83.3	83.3	75.0	66.7	75.0
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	50.0	25.0	25.0	25.0	50.0	37.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	66.7	80.0	80.0	66.7	66.7	72.7
Q3-32	SR	66.7	75.0	50.0	58.3	50.0	66.7
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	50.0	37.5	75.0	50.0	75.0	50.0
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	66.7	72.7	57.1	54.5	57.1	66.7
Show-o2	SR	50.0	33.3	33.3	33.3	33.3	50.0
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	75.0	100.0	100.0	100.0	100.0	75.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	57.1	50.0	50.0	50.0	50.0	57.1
VilU	SR	50.0	58.3	66.7	58.3	66.7	83.3
	FNR	50.0	100.0	100.0	75.0	100.0	0.0
	FPR	50.0	12.5	0.0	25.0	0.0	12.5
	R	50.0	0.0	0.0	25.0	0.0	100.0
	F1	40.0	0.0	0.0	28.6	0.0	88.9
LLaVA.N.Vid	SR	50.0	33.3	33.3	41.7	33.3	41.7
	FNR	0.0	25.0	0.0	0.0	0.0	0.0
	FPR	75.0	87.5	100.0	87.5	100.0	87.5
	R	100.0	75.0	100.0	100.0	100.0	100.0
	F1	57.1	42.9	50.0	53.3	50.0	53.3
Vid.LLaVA	SR	33.3	41.7	58.3	33.3	50.0	33.3
	FNR	0.0	25.0	25.0	25.0	25.0	0.0
	FPR	100.0	62.5	37.5	87.5	62.5	100.0
	R	100.0	75.0	75.0	75.0	75.0	100.0
	F1	50.0	50.0	60.0	42.9	50.0	50.0
Video-XL2	SR	50.0	58.3	50.0	33.3	41.7	50.0
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	75.0	62.5	75.0	100.0	87.5	75.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	57.1	61.5	57.1	50.0	53.3	57.1
M.avg	SR	55.0	55.8	50.8	48.3	50.0	58.3
	FNR	15.0	25.0	22.5	27.5	20.0	10.0
	FPR	60.0	52.5	61.3	63.7	65.0	56.2
	R	85.0	75.0	77.5	72.5	80.0	90.0
	F1	51.9	48.0	46.6	44.5	47.1	58.0

Table 7: Detailed detection performance on the Sora2 subset, categorized by cognitive scenarios.

Model	Indicator	None	Perception	Physical	Semantics	Reasoning	Culture
Human	SR	91.7	70.8	70.8	83.3	79.2	79.2
	FNR	12.5	56.2	62.5	31.2	37.5	50.0
	FPR	6.2	15.6	12.5	9.4	12.5	6.2
	R	87.5	43.8	37.5	68.8	62.5	50.0
	F1	87.5	50.0	46.2	73.3	66.7	61.5
Q2.5-3	SR	66.7	58.3	66.7	66.7	41.7	58.3
	FNR	100.0	100.0	50.0	100.0	50.0	75.0
	FPR	0.0	12.5	25.0	0.0	62.5	25.0
	R	0.0	0.0	50.0	0.0	50.0	25.0
	F1	0.0	0.0	50.0	0.0	36.4	28.6
Q3-2	SR	33.3	33.3	33.3	33.3	41.7	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	100.0	100.0	87.5	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	50.0	50.0	53.3	50.0
Q3-4	SR	33.3	50.0	50.0	25.0	33.3	41.7
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	100.0	75.0	75.0	100.0	100.0	87.5
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	50.0	57.1	57.1	40.0	50.0	53.3
Q3-8	SR	33.3	41.7	58.3	25.0	41.7	41.7
	FNR	0.0	0.0	0.0	50.0	0.0	0.0
	FPR	100.0	87.5	62.5	87.5	87.5	87.5
	R	100.0	100.0	100.0	50.0	100.0	100.0
	F1	50.0	53.3	61.5	30.8	53.3	53.3
Q3-32	SR	33.3	33.3	41.7	25.0	33.3	33.3
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	100.0	100.0	87.5	100.0	100.0	100.0
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	50.0	50.0	53.3	40.0	50.0	50.0
Show-o2	SR	33.3	33.3	33.3	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	100.0	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	50.0	50.0	50.0	50.0
VilU	SR	66.7	66.7	66.7	66.7	66.7	66.7
	FNR	50.0	100.0	75.0	100.0	100.0	25.0
	FPR	25.0	0.0	12.5	0.0	0.0	37.5
	R	50.0	0.0	25.0	0.0	0.0	75.0
	F1	50.0	0.0	33.3	0.0	0.0	60.0
LLaVA.N.Vid	SR	33.3	41.7	33.3	41.7	33.3	50.0
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	100.0	87.5	100.0	75.0	100.0	75.0
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	50.0	53.3	50.0	46.2	50.0	57.1
Vid.LLaVA	SR	16.7	33.3	41.7	33.3	75.0	41.7
	FNR	0.0	0.0	0.0	25.0	0.0	25.0
	FPR	100.0	87.5	75.0	87.5	37.5	62.5
	R	100.0	100.0	100.0	75.0	100.0	75.0
	F1	50.0	53.3	57.1	42.9	72.7	50.0
Video-XL2	SR	33.3	33.3	33.3	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	100.0	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	50.0	50.0	50.0	50.0
M.avg	SR	38.3	42.5	45.8	38.3	43.3	43.3
	FNR	15.0	20.0	12.5	35.0	15.0	12.5
	FPR	82.5	75.0	73.8	75.0	77.5	77.5
	R	85.0	80.0	87.5	65.0	85.0	87.5
	F1	45.0	41.7	51.2	35.0	46.6	50.2

Table 8: Detailed detection performance on the Wan2.2 subset, categorized by cognitive scenarios.

Model	Indicator	None	Perception	Physical	Semantics	Reasoning	Culture
Human	SR	62.5	68.8	72.9	66.7	85.4	72.9
	FNR	50.0	25.0	18.8	37.5	0.0	12.5
	FPR	31.2	34.4	31.2	31.2	21.9	34.4
	R	50.0	75.0	81.2	62.5	100.0	87.5
	F1	47.1	61.5	66.7	55.6	82.1	68.3
Q2.5-3	SR	66.7	66.7	58.3	50.0	41.7	66.7
	FNR	100.0	100.0	75.0	100.0	100.0	75.0
	FPR	0.0	0.0	25.0	25.0	37.5	12.5
	R	0.0	0.0	25.0	0.0	0.0	25.0
	F1	0.0	0.0	28.6	0.0	0.0	33.3
Q3-2	SR	33.3	33.3	41.7	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	87.5	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	53.3	50.0	50.0	50.0
Q3-4	SR	33.3	50.0	50.0	25.0	50.0	33.3
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	100.0	75.0	75.0	100.0	75.0	100.0
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	50.0	57.1	57.1	40.0	57.1	50.0
Q3-8	SR	66.7	75.0	41.7	33.3	75.0	50.0
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	50.0	37.5	87.5	87.5	37.5	75.0
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	66.7	72.7	53.3	42.9	72.7	57.1
Q3-32	SR	33.3	41.7	41.7	25.0	58.3	33.3
	FNR	0.0	0.0	0.0	25.0	0.0	0.0
	FPR	100.0	87.5	87.5	100.0	62.5	100.0
	R	100.0	100.0	100.0	75.0	100.0	100.0
	F1	50.0	53.3	53.3	40.0	61.5	50.0
Show-o2	SR	33.3	33.3	33.3	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	100.0	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	50.0	50.0	50.0	50.0
VilU	SR	50.0	66.7	66.7	41.7	58.3	83.3
	FNR	50.0	100.0	100.0	100.0	100.0	0.0
	FPR	50.0	0.0	0.0	37.5	12.5	25.0
	R	50.0	0.0	0.0	0.0	0.0	100.0
	F1	40.0	0.0	0.0	0.0	0.0	80.0
LLaVA.N.Vid	SR	50.0	33.3	33.3	33.3	33.3	41.7
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	75.0	100.0	100.0	100.0	100.0	75.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	57.1	50.0	50.0	50.0	50.0	57.1
Vid.LLaVA	SR	50.0	41.7	41.7	33.3	41.7	25.0
	FNR	0.0	25.0	25.0	50.0	25.0	0.0
	FPR	75.0	75.0	62.5	75.0	75.0	100.0
	R	100.0	75.0	75.0	50.0	75.0	100.0
	F1	57.1	46.2	50.0	33.3	46.2	50.0
Video-XL2	SR	33.3	33.3	41.7	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	87.5	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	53.3	50.0	50.0	50.0
M.avg	SR	45.0	47.5	45.0	34.2	45.8	43.3
	FNR	15.0	22.5	20.0	32.5	22.5	7.5
	FPR	75.0	67.5	71.2	82.5	70.0	78.8
	R	85.0	77.5	80.0	67.5	77.5	92.5
	F1	47.1	42.9	44.9	35.6	43.8	52.8

Table 9: Detailed detection performance on the Veo3.1-Fast subset, categorized by cognitive scenarios.

Model	Indicator	None	Perception	Physical	Semantics	Reasoning	Culture
Human	SR	50.0	64.6	66.7	68.8	70.8	83.3
	FNR	62.5	25.0	18.8	6.2	12.5	6.2
	FPR	43.8	40.6	40.6	43.8	37.5	21.9
	R	37.5	75.0	81.2	93.8	87.5	93.8
	F1	33.3	58.5	61.9	66.7	66.7	78.9
Q2.5-3	SR	33.3	66.7	58.3	58.3	58.3	58.3
	FNR	100.0	100.0	75.0	50.0	50.0	100.0
	FPR	50.0	0.0	25.0	37.5	37.5	12.5
	R	0.0	0.0	25.0	50.0	50.0	0.0
	F1	0.0	0.0	28.6	44.4	44.4	0.0
Q3-2	SR	33.3	33.3	41.7	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	87.5	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	53.3	50.0	50.0	50.0
Q3-4	SR	33.3	25.0	50.0	58.3	41.7	50.0
	FNR	0.0	25.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	75.0	62.5	87.5	75.0
	R	100.0	75.0	100.0	100.0	100.0	100.0
	F1	50.0	40.0	57.1	61.5	53.3	57.1
Q3-8	SR	33.3	25.0	41.7	58.3	41.7	50.0
	FNR	0.0	25.0	25.0	0.0	25.0	0.0
	FPR	100.0	100.0	75.0	62.5	75.0	75.0
	R	100.0	75.0	75.0	100.0	75.0	100.0
	F1	50.0	40.0	46.2	61.5	46.2	57.1
Q3-32	SR	33.3	33.3	50.0	50.0	50.0	41.7
	FNR	0.0	25.0	0.0	0.0	0.0	0.0
	FPR	100.0	87.5	75.0	75.0	75.0	87.5
	R	100.0	75.0	100.0	100.0	100.0	100.0
	F1	50.0	42.9	57.1	57.1	57.1	53.3
Show-o2	SR	33.3	33.3	33.3	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	100.0	100.0	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	50.0	50.0	50.0	50.0	50.0
VilU	SR	50.0	66.7	66.7	58.3	66.7	58.3
	FNR	100.0	100.0	100.0	50.0	100.0	0.0
	FPR	25.0	0.0	0.0	37.5	0.0	50.0
	R	0.0	0.0	0.0	50.0	0.0	100.0
	F1	0.0	0.0	0.0	44.4	0.0	66.7
LLaVA.N.Vid	SR	33.3	25.0	33.3	33.3	33.3	33.3
	FNR	0.0	25.0	25.0	0.0	0.0	25.0
	FPR	100.0	100.0	87.5	100.0	100.0	87.5
	R	100.0	75.0	75.0	100.0	100.0	75.0
	F1	50.0	40.0	42.9	50.0	50.0	42.9
Vid.LLaVA	SR	33.3	58.3	25.0	58.3	41.7	33.3
	FNR	0.0	25.0	50.0	0.0	0.0	0.0
	FPR	100.0	37.5	75.0	62.5	87.5	100.0
	R	100.0	75.0	50.0	100.0	100.0	100.0
	F1	50.0	60.0	33.3	61.5	53.3	50.0
Video-XL2	SR	33.3	41.7	41.7	33.3	33.3	33.3
	FNR	0.0	0.0	0.0	0.0	0.0	0.0
	FPR	100.0	87.5	87.5	100.0	100.0	100.0
	R	100.0	100.0	100.0	100.0	100.0	100.0
	F1	50.0	53.3	53.3	50.0	50.0	50.0
M.avg	SR	35.0	40.8	44.2	47.5	43.3	42.5
	FNR	20.0	32.5	27.5	10.0	17.5	12.5
	FPR	87.5	71.2	68.8	73.8	76.2	78.8
	R	80.0	67.5	72.5	90.0	82.5	87.5
	F1	40.0	37.6	42.2	53.1	45.4	47.7

C HUMAN EVALUATION DETAILS AND RESULTS

C.1 DETAILS OF HUMAN EVALUATION

In this study, we specifically recruited 24 human evaluators from diverse backgrounds to assess the authenticity of videos across different scenarios. The evaluators came from various age groups, genders, educational backgrounds, and cultural and geographical regions. Each evaluator independently judged the authenticity of the videos based on their personal cognition and experience. The diversity of the evaluators helps reduce bias, ensuring that the evaluation results are more representative and effectively reflect the perceptions of the general public regarding video authenticity.

Regarding the questionnaire design, human evaluators assess a subset of six generated videos in each scenario. In our questionnaire study, each participant evaluates 6 videos under the Zero-Knowledge Scenario and 60 videos with cognitive guidance. Each questionnaire therefore contains 66 videos. A participant typically requires about 40 minutes to complete a valid questionnaire, this duration follows the setup of the article on the questionnaire involving videos Jin et al. (2017). In total, 24 participants annotated 1584 video judgments according to the experimental protocol, resulting in approximately 960 minutes of video-based evaluation.

C.2 STATISTICAL ANALYSES OF HUMAN PERFORMANCE

Table 10 shows human detection performance across cognitive scenarios. Specifically, human evaluators assessed a subset of generated videos, including 36 videos under the Zero-Knowledge scenario and 72 videos under each cognitive-guidance scenario. We further report the exact numbers of video assessments in Table 10. For each scenario, We report the success rate (SR) alongside Wilson 95% confidence intervals (CI) to reflect estimation reliability across varying sample sizes. Statistical significance is assessed via two-sided binomial tests against chance-level performance ($p_0 = 0.5$). All scenarios satisfy $p < 0.001$, indicating that success rates are substantially above chance across every scenario and confirming that human evaluators possess genuine discriminative ability across different scenarios.

Table 10: Statistical Analyses of Human Performance across cognitive scenarios. **Total** is the number of video judgments collected in each scenario; **Correct** is the number of judgments that correctly identified the video as real or fake; **SR** = success rate (%); **CI** = Wilson 95% confidence interval (%); p = two-sided binomial test vs. chance ($p_0 = 0.5$). $^{***}p < 0.001$.

Scenario	Total	Correct	SR (%)	95% CI (%)	p vs. Chance
None	144	111	77.10	$[69.6,\ 83.2]$	$4.58 \times 10^{-11***}$
Perception	288	196	68.10	$[62.5,\ 73.2]$	$8.50 \times 10^{-10***}$
Physical	288	219	76.00	$[70.8,\ 80.6]$	$2.36 \times 10^{-19***}$
Semantics	288	218	75.70	$[70.4,\ 80.3]$	$7.46 \times 10^{-19***}$
Reasoning	288	238	82.60	$[77.8,\ 86.6]$	$1.69 \times 10^{-30***}$
Culture	288	237	82.30	$[77.5,\ 86.3]$	$7.96 \times 10^{-30***}$
Overall	**1,584**	**1,219**	**77.00**	$[\mathbf{74.8}, \mathbf{79.0}]$	$< \mathbf{0.001}^{***}$

D PROMPT CONSTRUCTION FOR EVALUATION

In this section, which corresponds to Section 4.2 and Section 4.4 of the main text, we provide the detailed prompts used to evaluate whether videos are AI-generated across various multimodal models. To evaluate the detection capabilities of multimodal large language models (MLLMs) under varying cognitive demands, we designed hierarchical prompt strategies corresponding to different scenario complexities.

The first strategy is the Basic Scenarios Setting, where the prompts incorporate detailed descriptions of cognitive information to provide a rich semantic basis for model reasoning and judgment. The second strategy is the Simple Scenario Setting, which retains only the most fundamental elements for judgment with brief guidance to test the core detection ability of the models. The third strategy is

the Complex Scenario Setting, where we introduce additional information regarding specific roles and styles along with highly detailed cognitive descriptions to challenge the models in more intricate contexts. By employing these different levels of prompting, we can comprehensively analyze how the depth of instructional guidance affects the detection performance and decision-making logic of MLLMs.

D.1 Basic Scenarios Setting

In the basic scenarios, we provide detailed descriptions of cognitive information tailored to different scenarios. These cognitive descriptions are designed to effectively guide the models in understanding and judging videos from various cognitive perspectives. By offering a rich semantic basis, these prompts enable the models to perform more reasoned evaluations beyond simple visual inspection. The specific evaluation prompts used in this setting are illustrated in Figure 12 and Figure 13.

Basic Scenario — Main Template

You will be given a video, please judge whether the given video is **Real** or **AI-Generated** through careful analysis.

{Cognitive Scenario Prompt}

IMPORTANT GENERAL PRINCIPLES

- Base analysis exclusively on the provided video content.
- Make a clear binary choice between 'Real Video' and 'AI Generated Video'.
- Do not output "Undetermined," "Cannot determine," or any non-decision response.
- Provide a concise explanation supporting your decision.

Judgement Example

Example 1:

Final decision: AI Generated Video

Reason: Unnatural fluid motion and frame drops were observed in the video, indicating AI-generated traces.

Example 2:

Final decision: Real Video

Reason: The screen is realistic and smooth, the characters are realistic, and the actions are natural.

Output Format

Final decision: Real Video *or* AI Generated Video

Reason: Briefly explain your reasoning for your choice in 1–3 sentences.

Figure 12: Overall evaluation prompts for different multimodal models. The "cognitive scenario prompt" is selected according to the prompts constructed for different cognitive scenarios.

Cognitive Scenario Prompt

Zero-Knowledge Scenario

Null

Perception and Senses Scenario

Real videos typically look natural and effortless, while AI can struggle with perfect consistency and physical realism.

Focus on **Visual Quality**: whether the video is clear, not blurry, and free of glitches or distortions.

Focus on **Color and Lighting Consistency**: whether colors and lighting look natural, consistent, and change logically.

Focus on **2D Layout**: whether objects are well-arranged and balanced in the frame.

Focus on **3D Depth**: whether the scene has a believable sense of depth and objects fit realistically in space.

Physical Plausibility Scenario

AI often shows errors when simulating the physical laws and interactions of the real world.

Focus on **Motion Realism**: whether motion looks smooth, lifelike, and not too rigid or floaty.

Focus on **Object Consistency**: whether objects keep the same appearance, texture, and size throughout.

Focus on **Object Interaction**: whether collisions, touches, or handling of objects feel realistic.

Focus on **Material Properties**: whether materials like water, cloth, or hair move as expected.

Focus on **Physical Laws**: whether gravity, momentum, and other basic physics feel real.

Semantics and Narrative Scenario

Real events have clear cause-and-effect. AI can create visually coherent but logically confusing sequences.

Focus on **Temporal Coherence**: whether the video frames are reasonable and the video feels coherent.

Focus on **Narrative Consistency**: whether the storyline maintains a coherent flow and makes sense logically throughout.

Focus on **Scene and Background Consistency**: whether the background stays stable and appropriate to the scene.

Focus on **Environment Interaction**: whether characters interact with surroundings in a believable way.

Focus on **Camera Control**: whether camera motions feel smooth, intentional, and physically possible.

Reasoning and Logic Scenario

AI-generated videos often have inconsistencies in logic or go against common sense.

Focus on **Causal Relation**: whether actions have clear reasons and understandable results.

Focus on **Character Attributes**: whether characters act in ways that fit their role or personality.

Focus on **Emotional and Expressive**: whether facial expressions and reactions match the situation.

Focus on **Common Sense**: whether the video follows everyday logic and doesn't break basic rules.

Focus on **Spatial Relationship**: whether the spatial relationships between objects are logical and consistent.

Social and Cultural Scenario

Real footage is grounded in specific, consistent details. AI can mix styles or get cultural specifics wrong.

Focus on **Cultural Element Authenticity**: whether cultural items, clothing, or customs are shown accurately.

Focus on **Social Background**: whether behaviors fit the social context shown.

Focus on **Atmosphere and Style**: whether the visual style, mood, and symbols remain uniform.

Figure 13: Cognitive information contained in different cognitive scenarios under the Basic Scenarios setting.

D.2 Simple Scenario Setting

In the simple scenario setting, we retain only the most fundamental elements for judgment and provide brief guidance information. These simplified prompts are designed to test the underlying capability of MLLMs to identify AI-generated content based solely on intuition and basic visual cues without detailed contextual support. In this way, we can evaluate the models' performance under extremely concise instructions. The specific evaluation prompts are presented in Figure 14 and 15.

Simple Scenario — Main Template

Given the following video, please determine if the video is **Real** or **AI-Generated**.

{Cognitive Scenario Prompt}

Output Format

Final decision: Real Video *or* AI Generated Video

Figure 14: Overall evaluation prompts for different multimodal models. The "cognitive scenario prompt" is selected according to the prompts constructed for different cognitive scenarios.

Cognitive Scenario Prompt

Zero-Knowledge Scenario

null

Perception and Senses Scenario

Carefully analyze from the Perception and Senses aspect.

Physical Plausibility Scenario

Carefully analyze from the Physics and Interaction aspect.

Semantics and Narrative Scenario

Carefully analyze from the Semantics and Narrative aspect.

Reasoning and Logic Scenario

Carefully analyze from the Reasoning and Logic aspect.

Social and Cultural Scenario

Carefully analyze from the Social and Culture aspect.

Figure 15: Cognitive information contained in different cognitive scenarios under the Simple Scenarios setting.

D.3 Complex Scenario Setting

In the complex scenario setting, we introduce additional information regarding roles, styles, and highly detailed cognitive descriptions. These intricate prompts are designed to construct a highly detailed contextual environment to challenge whether MLLMs can accurately extract key cognitive contradictions and perform logical reasoning when processing lengthy and complex information. In this way, we can observe the models' detection robustness when facing distracting information or high-dimensional detailed descriptions. The specific evaluation prompts are presented in Figure 16 and Figure 17.

Complex Scenario — Main Template

You are an **expert in AI-Generated Video Detection** *(role)*. Your analysis style is direct, concise, and **strictly fact-based** *(style)*. You will be given a video, please analyze **step by step** *(CoT)* whether the given video is **Real** or **AI-Generated**.

{Cognitive Scenario Prompt}

IMPORTANT GENERAL PRINCIPLES

- Base analysis exclusively on the provided video content.
- Make a clear binary choice between 'Real Video' and 'AI Generated Video'.
- Do not output "Undetermined," "Cannot determine," or any non-decision response.
- Provide a concise explanation supporting your decision.

Judgement Example

Example 1:

Final decision: AI Generated Video
Reason: Unnatural fluid motion and frame drops were observed in the video, indicating AI-generated traces.

Example 2:

Final decision: Real Video
Reason: The screen is realistic and smooth, the characters are realistic, and the actions are natural.

Output Format

Final decision: Real Video *or* AI Generated Video

Figure 16: Overall evaluation prompts for different multimodal models. The "cognitive scenario prompt" is selected according to the prompts constructed for different cognitive scenarios.

Cognitive Scenario Prompt

Zero-Knowledge Scenario

Null

Perception and Senses Scenario

Real videos typically look natural and effortless, whereas AI may struggle with achieving perfect consistency, especially in scenes with motion or complex lighting changes.

Focus on **Visual Quality**: Check whether the video is clear, sharp, and free from visual artifacts. Watch for any blurriness, pixelation, or unusual glitches that might indicate AI generation.

Focus on **Color and Lighting Consistency**: Observe if the colors in the video remain natural and consistent throughout, and ensure there are no abrupt changes or unrealistic lighting that might signal artificial creation.

Focus on **2D Layout**: Look at how objects and characters are placed within the frame. Check if the composition feels natural and if objects are well-arranged and proportionate.

Focus on **3D Depth**: Evaluate whether the scene has a realistic sense of depth and whether objects are positioned in a way that makes sense within the space. Watch for any awkward positioning or elements that seem flat or out of place.

Physical Plausibility Scenario

AI-generated videos often struggle to mimic the real physical world, such as how objects move or how gravity behaves.

Focus on **Motion Realism**: Watch how characters and objects move in the video. Smooth, natural movement is typical of real videos, while rigid or floaty motion may indicate AI creation. Make sure the motion flows seamlessly without sudden jerks or awkward pauses.

Focus on **Object Consistency**: Pay attention to whether objects stay consistent in size, shape, and texture throughout the video. Look for any visible changes or inconsistencies that may suggest AI generation.

Focus on **Object Interaction**: Watch how objects and characters interact with one another. Check if an object behaves unexpectedly during a collision, or if a character's hand passes through another object.

Focus on **Material Properties**: Observe how materials behave in the video. Water should flow naturally, cloth should move with the character's movements, and hair should sway in a realistic way.

Focus on **Physical Laws**: Look at how gravity, momentum, and other basic physical principles are treated. Ensure actions like falling objects, moving characters, or fluid dynamics follow the laws of nature.

Semantics and Narrative Scenario

In real-world videos, events unfold in a logical sequence with a clear relationship between actions and consequences. AI may create disjointed sequences where events feel out of order or disconnected.

Focus on **Temporal Coherence**: Ensure that the video flows naturally over time, with each scene following a reasonable progression. Check that the transitions between frames and sequences are consistent and reflect a realistic timeline.

Focus on **Narrative Consistency**: Watch for any changes in characters' roles, appearance, or behavior that break the flow of the story. Ensure that the narrative follows a clear path without contradictions or sudden shifts in plot.

Focus on **Scene and Background Consistency**: Check if the background remains stable and appropriate for the scene. If the setting changes without a logical reason or doesn't match the narrative, it may suggest AI generation.

Focus on **Environment Interaction**: Watch how characters interact with the environment. Ensure that their movements feel natural and appropriate for the scene. Awkward or ignored object interactions could indicate an AI-generated video.

Focus on **Camera Control**: Observe how the camera moves during the video. Smooth and purposeful camera movements help tell the story, while abrupt or jerky camera movements may be a sign of AI creation.

Reasoning and Logic Scenario

AI-generated videos often have actions or sequences that lack clear causes and effects, resulting in illogical or confusing storylines.

Focus on **Causal Relation**: Check if the events in the video have a clear cause-and-effect relationship. Ensure that actions lead to predictable results and avoid sequences that seem random or disconnected.

Focus on **Character Attributes**: Watch how characters behave in the video. Ensure that their actions align with their personalities and roles within the story. Inconsistencies in character behavior can indicate an AI-generated video.

Focus on **Emotional and Expressive**: Observe whether characters' facial expressions and emotional reactions are appropriate for the context of the scene. Inconsistent or exaggerated emotional responses can suggest AI generation.

Focus on **Common Sense**: Ensure that the video follows everyday logic and principles. Actions should be reasonable and realistic, without breaking any common sense rules or exhibiting out-of-place behaviors.

Focus on **Spatial Relationship**: Pay attention to how objects and characters are positioned in the scene. Ensure their spatial relationships make sense and follow natural principles such as gravity and perspective.

Social and Cultural Scenario

Cultural context is vital for distinguishing real footage from AI-generated content, which may misrepresent specific cultural elements or social norms.

Focus on **Cultural Element Authenticity**: Watch for signs of cultural accuracy in the video, such as clothing, language, and customs. Ensure these elements are presented correctly and in context with the story or setting.

Focus on **Social Background**: Evaluate the social behavior of characters within the scene. Ensure that their actions and behaviors align with cultural expectations and the social setting of the video.

Focus on **Atmosphere and Style**: Look at how the visual style, mood, and symbols are used in the video. Ensure that they remain consistent and reflect the appropriate cultural or social setting throughout.

Figure 17: Cognitive information contained in different cognitive scenarios under the Complex Scenarios setting.

E FACTORS BEHIND MODEL-BASED DETECTION PERFORMANCE

In this section, which corresponds to Section 4.4 of the main text, we provide the specific numerical results of detection performance for various multimodal models under the influence of different factors. Specifically, Appendix E.1 presents the main detection results under various conditions, while Appendix E.2 provides the detailed detection results across three scenarios. More detailed analysis is provided in the main text.

E.1 MAIN DETECTION RESULTS UNDER DIFFERENT FACTORS

The following content demonstrates the detection performance of humans and various Multi-modal Large Language Models (MLLMs). To evaluate detection performance of different model sizes and paradigms across various scenarios, we utilized a randomized subset of videos, including both real videos and AI-generated videos of high quality. We use Success Rate (SR) to measure detection performance. Detailed experimental results are shown in Table 11.

Table 11: Impact of Model Size and Paradigm on Model-Based Detection Performance.

Model	None	Perception	Physical	Semantics	Reasoning	Culture
Q2.5-3	54.4	56.9	56.7	58.4	49.3	58.1
Q3-2	34.4	33.6	33.6	33.6	34.4	34.1
Q3-4	35.6	37.8	36.7	38.2	37.1	39.3
Q3-8	41.1	43.3	44.4	47.1	47.1	47.8
Q3-32	54.4	55.6	54.4	57.8	58.0	59.3
Show-o2	40.0	36.7	35.1	37.6	38.2	35.2
VilU	30.0	37.5	62.2	37.1	63.1	37.8
LLaVA.N.Vid	38.9	55.6	42.9	51.3	44.9	53.0
Vid.LLaVA	36.7	40.3	39.3	42.7	40.9	47.8
Video-XL2	35.6	35.3	34.9	34.9	36.0	35.6
M.avg	40.1	43.3	44.0	43.9	44.9	44.8

E.2 MODEL-BASED DETECTION PERFORMANCE ACROSS DIFFERENT SCENARIO COMPLEXITIES

To analyze model-based detection performance across different scenario complexities, we randomly select videos from the CogDetect dataset, covering both AI-generated and real videos of all categories. Tables 12, 13, and 14 present the Success Rate (SR) of different multimodal models under the Basic, Simple, and Complex scenario settings. The final row of the table represents the average detection performance across all models.

Table 12: Detection performance under Basic Scenario.

Model	None	Perception	Physical	Semantics	Reasoning	Culture
Q2.5-3B	64.3	62.5	59.3	63.6	62.1	65.5
Q3-2B	53.6	46.4	52.9	51.4	54.3	51.2
Q3-4B	64.3	67.0	67.1	70.0	62.1	66.7
Q3-8B	71.4	69.6	67.1	64.3	65.7	69.0
Q3-32B	67.9	67.9	67.9	67.9	71.4	67.9
Show-o2	39.3	33.9	38.6	42.9	45.0	47.6
VilU	67.9	60.7	65.7	58.6	65.7	58.3
LLaVA.N.Vid	53.6	40.2	32.1	39.3	37.1	61.9
Vid.LLaVA	42.9	53.6	51.4	47.1	60.7	32.1
Video-XL2	53.6	53.6	53.6	53.6	53.6	53.6
M.avg	57.9	55.5	55.6	55.9	57.8	57.4

Table 13: Detection performance under Simple Scenario.

Model	None	Perception	Physical	Semantics	Reasoning	Culture
Q2.5-3B	64.3	57.1	64.3	60.7	57.1	60.7
Q3-2B	35.7	35.7	35.7	35.7	39.3	35.7
Q3-4B	71.4	67.9	67.9	71.4	71.4	67.9
Q3-8B	71.4	71.4	71.4	71.4	71.4	67.9
Q3-32B	67.9	71.4	67.9	67.9	71.4	67.9
Show-o2	32.1	32.1	32.1	32.1	32.1	32.1
VilU	50.0	46.4	53.6	42.9	46.4	46.4
LLaVA.N.Vid	35.7	35.7	25.0	28.6	32.1	39.3
Vid.LLaVA	25.0	21.4	21.4	25.0	21.4	25.0
Video-XL2	53.6	53.6	50.0	32.1	32.1	53.6
M.avg	50.7	49.3	48.9	46.8	47.5	49.6

Table 14: Detection performance under Complex Scenario.

Model	None	Perception	Physical	Semantics	Reasoning	Culture
Q2.5-3B	67.9	67.9	67.9	67.9	67.9	67.9
Q3-2B	75.0	77.7	72.1	72.9	69.3	75.0
Q3-4B	75.0	74.1	75.7	74.3	75.0	75.0
Q3-8B	71.4	71.4	69.3	70.0	70.7	71.4
Q3-32B	67.9	65.2	66.4	67.1	70.7	66.7
Show-o2	32.1	32.1	32.1	32.1	32.1	32.1
VilU	42.9	39.3	32.9	37.9	50.0	45.2
LLaVA.N.Vid	60.7	48.2	34.3	43.6	44.3	38.1
Vid.LLaVA	17.9	19.6	10.7	20.7	16.4	15.5
Video-XL2	57.1	55.4	60.0	57.1	53.6	53.6
M.avg	56.8	55.1	52.1	54.4	55.0	54.1

F Detailed Description of Comparative Experiments

To further analyze the comparative experiments mentioned in the main text Section 5.2, we provide the detailed configurations in this Section. All evaluations are conducted on the GenVideo-val dataset, which is a standard benchmark for AI-generated video detection. This dataset contains diverse samples from various generation models such as Sora, I2VGen-XL, and ZeroScope, along with real videos from MSR-VTT and Kinetics-400. To ensure a fair comparison with baseline methods like NSG-VD and DeMamba, we follow the same preprocessing steps and use the Success Rate (SR), Recall and AUC score as evaluation metrics. The evolved promp is illustrated in Figure 18.

The implementation of our framework differs significantly from existing supervised models. While models require training on large datasets, our method is entirely training-free. We use only 7 generated videos as initial seeds and optimize the instructions through an iterative process guided by five different cognitive scenarios. Our evolutionary process naturally stops after optimizing the instruction for information extraction on 7 different videos across 5 distinct scenarios. This approach does not involve any parameter updates or gradient descent, which avoids the high computational costs and data requirements associated with traditional training.

The experimental results show that our training-free method achieves a competitive Success Rate. Although there is a small performance gap compared to some fully trained models on this specific dataset, our approach offers distinct advantages in practical applications. Because it does not rely on specific training artifacts, it demonstrates stronger generalization on unseen video data and new generative models. This flexibility allows our method to be deployed easily in real-world environments without the need for constant retraining as new generation technologies emerge.

Cross-dataset generalization. We further evaluate our method on the GenVidBench and AIGVD-Bench datasets, with the results reported in Table 15. The results show that the model's average SR improves from 54.33% to 74.62% on GenVidBench, with an increase of 20.29 percentage points, and from 52.75% to 79.14% on AIGVDBench, with an increase of 26.39 percentage points, providing additional evidence of cross-dataset generalization.

G Understanding of the Selected Cognitive Scenarios

G.1 Explanation of Scenario Selection

Following the typical visual cognition literature Summerfield & Egner (2009); Cavanagh (2011); Alvarez (2011); Humphreys & Bruce (2021); Petrusel et al. (2017), we select five scenarios that cover the full hierarchy of human cognition. They range from low-level perception to high-level reasoning and provide a comprehensive perspective for detecting AI-generated videos. The results of

The Evolved Prompt

You will be given a video, please judge whether the given video is **Real** or **AI-Generated** through careful analysis.

IMPORTANT Tips:

1. **Lighting Consistency:** Check for sudden changes in lighting that do not align with natural light patterns. This can indicate artificial lighting sources or inconsistent exposure settings.
2. **Texture and Detail:** Look for areas where textures appear unnatural or overly smooth. This could be due to the use of filters or post-production effects.
3. **Color Saturation:** Examine if the color saturation is too high or too low, which might suggest an attempt to correct for a specific color temperature or filter.
4. **Hand Movements and Interaction:** Pay close attention to the precision and fluidity of hand movements. Inconsistent or unnatural hand movements can be a sign of AI-generated content.
5. **Consistency in Appearance of Gloves:** Ensure that the appearance of gloves remains consistent across different frames. Any sudden changes in glove appearance could indicate AI generation.
6. **Atmosphere and Style:** Evaluate if the overall atmosphere and style of the video remain uniform throughout. Discontinuities in the visual style can hint at AI-generated content.
7. **Behavioral Consistency:** Confirm that the behaviors of characters or objects fit the expected social context. Inconsistencies in behavior can be a strong indicator of AI-generated content.
8. **Artificial Hand Gestures:** Watch for any sudden or unnatural hand gestures that do not match typical human movements. This can be a clear indicator of AI-generated content.
9. **Unnatural Object Interactions:** Examine if there are any interactions between objects that seem too perfect or unrealistic. This can suggest the use of AI-generated content.
10. **Inconsistent Backgrounds:** Check for backgrounds that change abruptly or do not match the expected environment. This can indicate the use of AI-generated content.
11. **Unrealistic Hand Movements:** Look for hand movements that are overly precise, jerky, or lack natural fluidity, which can be a sign of AI-generated content.
12. **Unnatural Finger Gestures:** Pay attention to finger gestures that appear too rigid or unnatural, which can be a clear indicator of AI-generated content.
13. **Discontinuous Backgrounds:** Ensure that the background does not change abruptly or does not match the expected environment. This can indicate the use of AI-generated content.
14. **Unrealistic Hand Postures:** Examine if the hand postures are too rigid, unnatural, or lack the subtlety of real human hand movements, which can be a sign of AI-generated content.
15. **Unnatural Hand Positioning:** Watch for hand positions that are too static or unnatural, which can be a clear indicator of AI-generated content.
16. **Unrealistic Camera Movement:** Check for sudden or unnatural camera movements that do not match typical human actions. This can be a clear indicator of AI-generated content.
17. **Unnatural Lighting Effects:** Examine if there are any lighting effects that seem too perfect or unrealistic, such as overly bright spots or shadows. This can suggest the use of AI-generated content.
18. **Unnatural Sound Effects:** Pay attention to sound effects that do not match the visual content. Sudden or unnatural sound effects can be a sign of AI-generated content.
19. **Unrealistic Facial Expressions:** Examine if facial expressions are too rigid, unnatural, or lack the subtlety of real human expressions. This can be a clear indicator of AI-generated content.
20. **Unnatural Object Transformations:** Watch for object transformations that seem too perfect or unrealistic, such as sudden changes in shape or size. This can suggest the use of AI-generated content.
21. **Unnatural Hand Posture Dynamics:** Look for hand postures that are too rigid, unnatural, or lack the subtlety of real human hand movements, which can be a sign of AI-generated content.
22. **Unnatural Finger Gesture Dynamics:** Pay attention to finger gestures that appear too rigid or unnatural, which can be a clear indicator of AI-generated content.
23. **Discontinuous Background Dynamics:** Ensure that the background does not change abruptly or does not match the expected environment. This can indicate the use of AI-generated content.
24. **Unrealistic Hand Movement Dynamics:** Look for hand movements that are overly precise, jerky, or lack natural fluidity, which can be a sign of AI-generated content.
25. **Unnatural Hand Positioning Dynamics:** Watch for hand positions that are too static or unnatural, which can be a clear indicator of AI-generated content.
26. **Unrealistic Camera Movement Dynamics:** Check for sudden or unnatural camera movements that do not match typical human actions. This can be a clear indicator of AI-generated content.
27. **Unnatural Lighting Effect Dynamics:** Examine if there are any lighting effects that seem too perfect or unrealistic, such as overly bright spots or shadows. This can suggest the use of AI-generated content.

IMPORTANT GENERAL PRINCIPLES

- Base analysis exclusively on the provided video content.
- Make a clear binary choice between 'Real Video' and 'AI Generated Video'.
- Do not output "Undetermined," "Cannot determine," or any non-decision response.
- Provide a concise explanation supporting your decision.

Judgement Example

Example 1:

Final decision: AI Generated Video
Reason: The video shows 'identity morphing' where a hand-held object changes shape when obscured, and the background text consists of nonsensical gibberish symbols.

Example 2:

Final decision: Real Video
Reason: The physical interactions between the tool and the material show realistic resistance, and the lighting/shadows maintain perfect spatial logic throughout the movement.

Output Format

Final decision: Real Video *or* AI Generated Video
Reason: Briefly explain your reasoning for your choice in 1-3 sentences.

Figure 18: An example of the evolved prompt produced by our instruction evolution framework.

Table 15: **Cross-dataset generalization results.** Comparison between the base prompt and the evolved instruction of the proposed framework. Both instructions are evaluated on three benchmarks, including GenVideo-val, GenVidBench, and AIGVDBench. We report detection SR on each benchmark. Our evolved instruction consistently outperforms the base instruction and demonstrates strong cross-dataset generalization.

Benchmark	**Success Rate (%)**		**Δ (pts)**
	Base Prompt	**Evolution(ours)**	
GenVideo-val	52.70	77.09	+24.39
GenVidBench	54.33	74.62	+20.29
AIGVDBench	52.75	**79.14**	+26.39

human performance across the different cognitive scenarios indicate that the constructed hierarchical scenarios align with human cognitive habits and are both well grounded and meaningful. Furthermore, the results of model performance under each cognitive scenario show that the detection performance of capable models improves as the level of cognitive information embodied in the scenario increases, suggesting that richer cognitive cues contribute to more reliable detection.

G.2 Analyses of Mixed-Scenario Guidance

To further examine the guiding role of scenario mixtures in detection, we conduct two additional experiments and report the average performance of several representative models. Specifically, to assess simple mixed-scenario guidance, we progressively combine the five scenarios along the cognitive hierarchy, from low-level perception to high-level social and cultural reasoning, and present the results in Table 16(a). We then apply the instruction self-evolution framework under different mixed-scenario configurations to investigate how scenario composition affects the evolution process and detection performance, with the results reported in Table 16(b).

Simple mixed-scenario guidance. Using the Zero-Knowledge Scenario as the reference, we observe that each individual cognitive scenario improves detection accuracy, and the accuracy shows an increasing trend with the enhancement of cognitive guidance levels. In contrast, the accuracy of mixed scenarios all decreased. This suggests that a single scenario can help the model understand videos to varying degrees. But when scenarios are simply mixed, the model cannot accurately extract sufficient effective information from the mixed scenarios. Instead, the introduction of excessive complex information may harm the accuracy of detection. We also reached the same conclusion through experiments on scenario complexity.

Effect of scenario composition on the evolution process and detection performance. The independently evolved instructions generally achieve higher accuracy under higher-level cognitive scenarios, further supporting the rationale and effectiveness of our scenario design. As more evolved scenario-specific instructions are integrated, performance consistently improves, and reaches its best result when all five scenarios are included. This demonstrates that our framework can effectively exploit and integrate information from different cognitive scenarios while reducing the redundancy and interference in the simple mixed-scenario.

Table 16: Analyses of mixed-scenario guidance. (a) Simple mixed-scenario guidance: each individual scenario improves SR, with larger gains at higher cognitive levels, whereas simply mixing scenarios decreases accuracy. (b) The instruction self-evolution framework under different guidance configurations: individually evolved instructions perform better under higher-level scenarios, and performance improves consistently as more evolved instructions are integrated, achieving the best result with all five scenarios.

(a) Simple mixed-scenario guidance

Mix of Scenarios	SR (%)	FNR (%)	FPR (%)	ΔSR vs. None
None	**46.36**	**4.24**	**78.33**	**Reference**
Perception only	47.58	5.76	75.76	+1.22
Physical only	47.63	10.76	73.18	+1.27
Semantics only	49.19	8.64	71.89	+2.83
Reasoning only	47.47	6.82	75.38	+1.11
Social and Cultural only	48.38	5.91	74.47	+2.02
Perception + Physical	45.35	7.27	78.33	−1.01
Perception + Physical + Semantics	46.26	6.21	77.50	−0.10
Perception + Physical + Semantics + Reasoning	45.66	7.12	77.95	−0.70
Perception + Physical + Semantics + Reasoning + Social and Cultural	46.21	8.64	76.36	−0.15

(b) Self-evolution under different guidance configurations

Guidance Configuration	Base SR (%)	Evolved SR (%)	ΔSR
Perception only	52.71	54.11	+1.40
Physical only	52.71	53.80	+1.09
Semantics only	52.71	54.07	+1.36
Reasoning only	52.71	56.23	+3.52
Social and Cultural only	52.71	59.75	+7.04
Perception + Physical	52.71	55.29	+2.58
Perception + Physical + Semantics	52.71	61.52	+8.81
Perception + Physical + Semantics + Reasoning	52.71	71.65	+18.94
Perception + Physical + Semantics + Reasoning + Social and Cultural	52.71	**77.09**	+24.39