Beyond accuracy: Tracking more like Human via Visual Search

Dailing Zhang[1,2] **Shiyu Hu**[5] **Xiaokun Feng**[1,2]
Xuchen Li[1,2] **Meiqi Wu**[3] **Jing Zhang**[2] **Kaiqi Huang**[1,2,4]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences
[3]School of Computer Science and Technology, University of Chinese Academy of Sciences
[4]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
[5]School of Physical and Mathematical Sciences, Nanyang Technological University
zhangdailing2023@ia.ac.cn, shiyu.hu@ntu.edu.sg, fengxiaokun2022@ia.ac.cn,
wumeiqi18@mails.ucas.ac.cn, {lixuchen2024, jing_zhang, kqhuang}@ia.ac.cn

Abstract

Human visual search ability enables efficient and accurate tracking of an arbitrary moving target, which is a significant research interest in cognitive neuroscience. The recently proposed Central-Peripheral Dichotomy (CPD) theory sheds light on how humans effectively process visual information and track moving targets in complex environments. However, existing visual object tracking algorithms still fall short of matching human performance in maintaining tracking over time, particularly in complex scenarios requiring robust visual search skills. These scenarios often involve **S**patio-**T**emporal **D**iscontinuities (*i.e.*, *STDChallenge*), prevalent in long-term tracking and global instance tracking. To address this issue, we conduct research from a human-like modeling perspective: (1) Inspired by the CPD, we propose a new tracker named **CPDTrack** to achieve human-like visual search ability. The central vision of CPDTrack leverages the spatio-temporal continuity of videos to introduce priors and enhance localization precision, while the peripheral vision improves global awareness and detects object movements. (2) To further evaluate and analyze *STDChallenge*, we create the ***STDChallenge Benchmark***. Besides, by incorporating human subjects, we establish a human baseline, creating a high-quality environment specifically designed to assess trackers' visual search abilities in videos across *STDChallenge*. (3) Our extensive experiments demonstrate that the proposed CPDTrack not only achieves state-of-the-art (SOTA) performance in this challenge but also narrows the behavioral differences with humans. Additionally, CPDTrack exhibits strong generalizability across various challenging benchmarks. In summary, our research underscores the importance of human-like modeling and offers strategic insights for advancing intelligent visual target tracking. Code and models are available at `https://github.com/ZhangDailing8/CPDTrack`.

1 Introduction

In the real world, humans excel at locating an arbitrary moving target within complex backgrounds and can resume tracking it even after temporary loss. This enduring question in cognitive neuroscience has recently been explained to some extent by the Central-Peripheral Dichotomy (CPD) theory, which suggests that the human eyes processes all visual inputs by sorting them into central and peripheral visions [1, 2]. Central vision, decoded and understood by higher brain areas, focuses on interpreting details and minimizing distractions; while peripheral vision, processed in the primary visual cortex (V1), swiftly detects dynamic changes.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0086

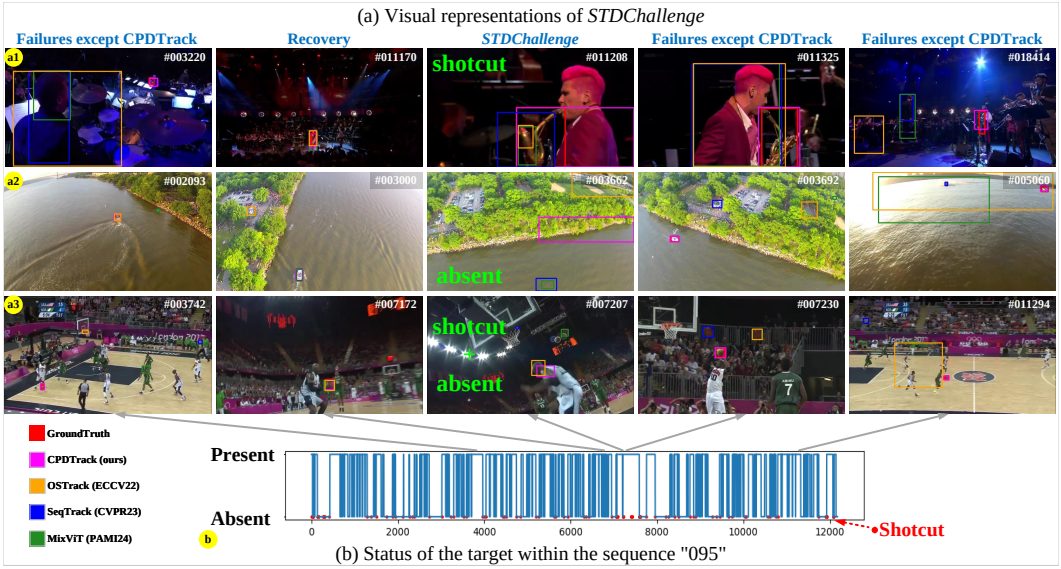

Figure 1: Illustration of the *STDChallenge*, depicting the *absent* of targets and *shotcut*. *STDChallenge* is quite challenging, but CPDTrack can maintain robust tracking performance, demonstrating stronger visual search abilities compared to other trackers. (a) In the first column, most trackers fail; in the second column, they can recover; in the third column, a *STDChallenge* occurs; in the fourth column, all trackers except for CPDTrack fail; and in the fifth column, this remains the case, meaning trackers' limited recovery. (b) shows the status of the target within the sequence of "095" from VideoCube [4], and red dots means *shotcut*.

Meanwhile, in computer vision research, advancements in single object tracking (SOT) are increasingly closing the gap between human dynamic visual ability (DVA) and proxy tasks [3, 4, 5]. In 2013, the short-term tracking (STT) task and related benchmarks were proposed [6, 7, 8]. However, the implicit assumption of continuous motion simplifies this task to continuously locating the target in short videos of tens of seconds, which is far from human DVA. Drawing on human visual search skills, the VOT community is broadening its scope by incorporating long-term tracking (LTT) [9, 10], which includes target *absent* (*i.e.*, the moving target can disappear and reappear in tracking), and global instance tracking (GIT) [4], which involves *shotcut* (*i.e.*, the video sequence may include different viewpoints and scenes). We refer to these challenges collectively as **spatio-temporal discontinuity (*STDChallenge*)**, which demands enhanced visual search abilities from trackers, as shown in Fig. 1.

Unfortunately, mainstream machines have not kept pace with the aforementioned expansion of task definitions to achieve human performance. Most mainstream trackers [11, 12, 13, 14, 15, 16], influenced by the definition of STT, depend on the spatio-temporal continuity of the target. These trackers typically perform local cropping based on the previous frame's results within the motion model, a module for calculating the search region on the current frame, as shown in Fig. 2(b-1). Obviously, this modeling mechanism only simulates human central vision, making it far from replicating human visual search ability, especially in scenarios where the *STDChallenge* increases significantly, as shown in Fig. 6(a). Several studies [17, 18] have also demonstrated that *STDChallenge* presents a significant challenge for trackers. Thus, a natural question is: **what causes trackers to overlook the importance of the *STDChallenge* in their modeling process**?

Some possible reasons from the benchmark perspective may answer the above question: **Limitation of datasets.** Although the definition of proxy tasks has been moving closer to human DVA (STT $\rightarrow$ LTT $\rightarrow$ GIT), most datasets are still influenced by the initial task characteristics (*i.e.*, *single-target*, *model-free*, *causal-trackers*, *single-camera*, and *short-term*) in data collection process [19, 20, 21, 9], as in Fig. 10, resulting in the importance of the *STDChallenge* being overlooked. Consequently, trackers can easily achieve good performance on most datasets without possessing sufficient visual search ability to handle the *STDChallenge*. **Limitation of evaluation.** Researchers contend that integrating insights from human DVA could address *STDChallenge* and enhance the real-world robustness of machines [22, 3, 23]. However, the substantial differences in methodologies and experimental

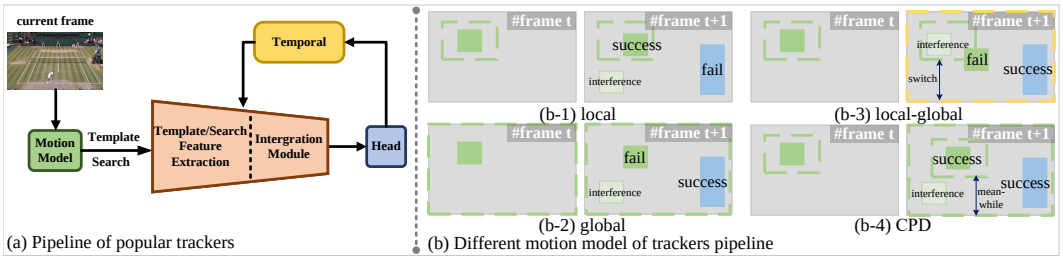

Figure 2: Comparison of tracking pipeline. Our CPDTrack differs from previous trackers in motion model. (a) The core components of mainstream tracking frameworks consist of a motion model, feature extraction, and a temporal module. (b-1) local tracker, which tracks targets in local areas and has difficulty recovering after failure; (b-2) global tracker, which tracks targets globally in the current frame, is susceptible to background interference and has lower efficiency; (b-3) local-global tracker, which can switch between the above two, depending on the performance of the local tracker; (b-4) our CPDTrack, which can model both central and peripheral information simultaneously.

approaches between computer science and cognitive science make it difficult to determine if current technological advancements are truly bridging the gap between SOT machines and human DVA. Consequently, there is a notable absence of mechanisms to assess the true visual intelligence of trackers.

In summary, existing trackers lack consideration of human visual search ability in their modeling mechanisms, making it difficult to cope with the *STDChallenge*. This limitation is not only due to the algorithm design itself but also involves the limitations of datasets and evaluation. Therefore, this work aims to address these aspects (algorithms and benchmark) and verifies the relevant results through experimental analysis.

A new algorithm CPDTrack (Section 3). Our analysis of the tracker pipeline and understanding of the CPD suggests that mainstream motion models, which crop local regions from the current frame, as shown in Fig. 2, restrict the tracker's ability to locate moving targets. Drawing on the distinction between central and peripheral vision in CPD, we propose a new tracker named CPDTrack to divide the current frame accordingly. Central vision captures detailed information, leveraging the video's spatio-temporal continuity to introduce priors, while peripheral vision manages global information to enhance overall scene understanding, as shown in Fig. 3. CPDTrack also models the information query, achieving top-down control based on cognitive outcomes.

A new *STDChallenge Benchmark* with Visual Turing Test (Section 4). To address the shortcomings of existing benchmarks environments, we introduce a specialized challenge environment named *STDChallenge Benchmark* to assess the visual search abilities of both machines and humans in tracking an arbitrary moving target. This environment comprises sequences that represent the *STDChallenge*, carefully selected from both LTT and GIT benchmarks. By sampling from a variety of benchmarks, we ensure a reduction in dataset-specific biases. Simultaneously, we employ the Visual Turing Test (VTT) [24, 25] on the *STDChallenge Benchmark* to assess the disparity in intelligence between trackers and human DVA. Specifically, we examine error consistency [26] to compare the performance of humans and machines, and explore behavioral differences among machines with varying architectures and parameters in dynamic visual tasks. Building on these results, we conduct a thorough and detailed analysis to reveal the impact of the *STDChallenge* on trackers.

Comprehensive and integral experimental analyses (Section 5). Finally, analysis results from extensive experimental settings show that, in the *STDChallenge*, CPDTrack not only achieved state-of-the-art (SOTA) performance, getting 1.7% superior N-PRE and 1.4% PRE score on *STDChallenge Benchmark* compared to the second algorithm, as shown in Table 1, but also demonstrated remarkable alignment with human DVA, as shown in Fig. 4. Furthermore, it performed well across various tracking benchmarks, particularly excelling in difficult benchmarks [4, 9, 27, 28].

Contributions. Our research explores the visual search ability of humans and trackers through detailed analyses based on algorithmic strategies and benchmark evaluations.

https://doi.org/10.52202/079017-0086

2 Related Work

Tracking Benchmark. In 2013, the VOT competition characterized SOT using five key terms: *single-target*, *model-free*, *causal-trackers*, *single-camera* and *short-term*. The initial three terms serve to differentiate SOT from other visual tasks [29, 30, 31, 32], whereas the last two, *single-camera* and *short-term*, were introduced to simplify early-stage research [6, 7, 19, 21]. Since 2018, several researchers have moved beyond the *short-term* limitation to embrace LTT [33, 9, 34], with the VOT setting a new criterion that tasks permitting complete target disappearance qualify as long-term [10]. More recently, the introduction of GIT [4] has eliminated the *single-camera* restriction, thus offering a more realistic simulation of the world. *STDChallenge Benchmark* is dedicated to examining the distinctions between LTT and GIT compared to STT, and integrates human DVA to steer algorithmic advances.

Visual Trackers. Deep learning trackers are typically categorized by motion models into local and global trackers, as shown in Fig. 2. Local trackers, which are most popular method, crop a search region in current frame [11, 12, 13, 15, 35, 16, 36]. These trackers continuously improved by advancements in feature extractor and temporal analysis, have reached SOTA performance on various benchmarks [9, 19, 21]. To tackle *STDChallenge*, it is common to switch to a global detector when local trackers fail, though this transition depends on the local tracker's performance [37, 38, 39, 40, 41]. Local trackers often struggle with proactive global awareness, reducing their robustness against *STDChallenge*. Global trackers, conversely, aim to identify targets throughout the entire image, raising computational demands and easy to be confused by similar objects in the background. Due to these limitations, they are less commonly employed [42, 43]. CPDTrack is designed to model CPD by combining local and global information, thereby enabling precise and efficient global modeling and offering insights for addressing the *STDChallenge*. For a detailed analysis of these trackers, see Appendix D.

Visual Turing Test. Traditional evaluation methods, focused on "machine versus machine" comparisons, struggle to comprehensively assess the development of visual intelligence as they rely on predefined metrics in experimental settings. The VTT, at the intersection of turing tests and computer vision, is gaining attention for its "human versus machine" approach, which uses human visual capabilities as a baseline to fully evaluate machine effectiveness [26, 44, 45, 46, 26, 25, 24]. This paradigm has been applied in SOT that simulate human DVA: GIT[4] demonstrates that humans can effectively resolve *STDChallenge*, and PathTracker [22, 47] has developed a circuit model to explain human decision-making processes. We are going to explore the visual search ability of humans and machines when tracking an arbitrary moving target.

3 Solving the *STDChallenge*

This section presents the proposed CPDTrack method in detail. We introduce CPD motion model, which divides global information into central and peripheral vision to optimize feature extraction. Additionally, we have developed a temporal clues module that serves as the information query component within CPD theory, facilitating top-down control mechanisms.

3.1 CPD Motion Model

A crucial challenge is balancing the maintenance of local resolution and computational efficiency during the extraction of global information. To address this, we incorporate insights from cognitive neuroscience, specifically the fovea concept [48] and CPD, segmenting the current frame into central and peripheral. Their computation is divided into the following steps:

Determine the position. For the central vision, as shown in Fig. 3, we use the tracking result $b_{t-1} = (b^{cx}_{t-1}, b^{cy}_{t-1}, w_{t-1}, h_{t-1})$ on the previous frame to determine its position.

Central vision. To simulate the decay in visual sensitivity characteristic of the fovea, we employ gaussian distribution model—a method that has found widespread application in vision science [48, 49]. Visual sensitivity is defined as:

$$S(x) = S_0 \cdot exp(-\frac{\left(x - b^{cx}_{t-1}\right)^2}{2 \cdot \sigma^2}), \tag{1}$$

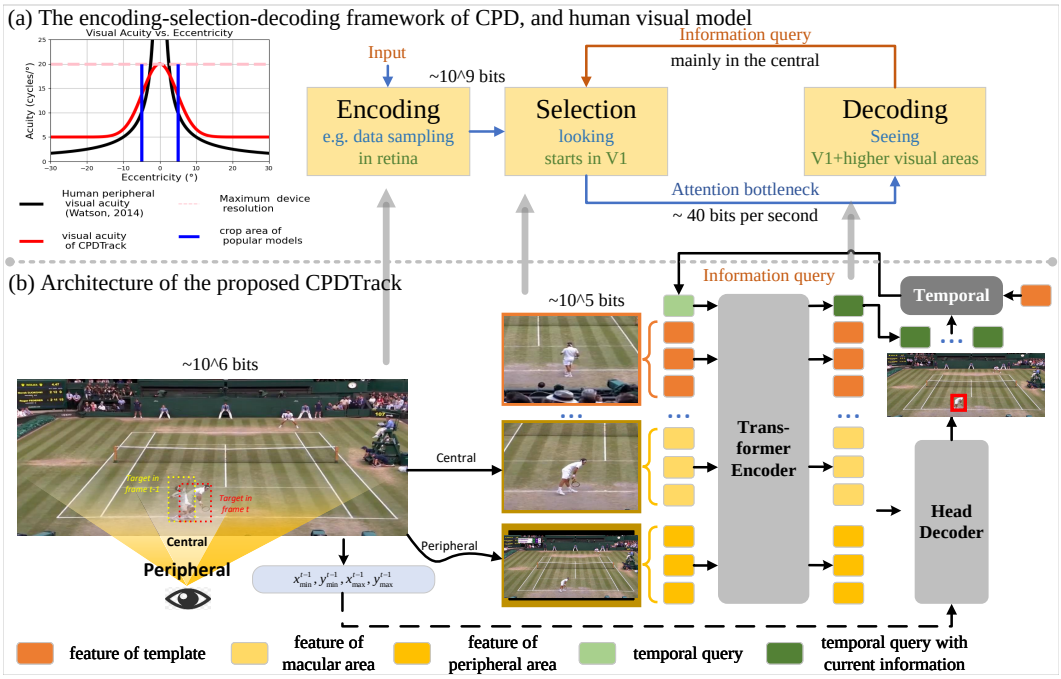

Figure 3: The overall architecture of CPDTrack, referencing the latest one-stream trackers, models CPD. (a) The encoding-selection-decoding framework of CPD. **Visual selection** is the process of choosing the information to focus on, and **visual decoding** is the process of deeply understanding the selected features and information to make cognitive decisions. On the left is a mathematical model of visual selection, where we align human visual information processing with the resolution of devices and machines. (b) Architecture of the proposed CPDTrack: The original frame is treated as encoded visual data. We apply the acuity model described in (a) for visual selecting, which can be modulated by **information query** from temporal. A transformer is employed to replicate the decoding processes occurring in higher brain regions, facilitating complex cognitive tasks. The grey arrows running through both parts highlight the correspondence between the two parts.

where $S(x)$ is sensitivity in x and S_0 is the sensitivity in central. we set the $\sigma = \frac{W}{6}$ based on the overall size (H, W) of the image and the 3-sigma rule. This approach ensures that the cumulative visual sensitivity to current frame exceeds 99.7%, effectively covering the vast majority of the visual field.

Sensitivity of central. We determine the sensitivity levels within this region using the values from b_{t-1}. Integrating sensitivity as:

$$sens_{x-1} = \int_{b_{t-1}^{x_1}}^{b_{t-1}^{x_2}} S(x)dx, \tag{2}$$

where $b_{t-1}^{x_1} = b_{t-1}^{cx} - \frac{w_{t-1}}{2}$ is the top-left coordinate of b_{t-1}, and $b_{t-1}^{x_2} = b_{t-1}^{cx} + \frac{w_{t-1}}{2}$ is the bottom-right coordinate of b_{t-1}.

Image crop resizing. As the central region is resized to a constant size, thus its width $w_{t-1}^e \propto \frac{1}{sens_x}$ and $w_{t-1}^e \propto w_{t-1}$. Therefore, w_{t-1}^e is determined to be:

$$w_{t-1}^e = \frac{w_{t-1}}{sens_{x-1}} = \mathcal{S}\frac{w_{t-1}}{2\Phi(\frac{3w_{t-1}}{W}) - 1}, \tag{3}$$

where $\mathcal{S}$ is a constant factor that can be adjusted, and Φ is the cumulative density function of the standard normal distribution. Then, we get central region $m_t = (b_{t-1}^{cx}, b_{t-1}^{cy}, w_{t-1}^e, h_{t-1}^e)$.

As the lower limit of the bounding box is set, typically at 10 pixels in tracking, equation (3) will not a division by zero error in practice.

https://doi.org/10.52202/079017-0086

We prove the range w_{t-1}^e in equation (3): First, take the derivative with respect to $w_{t-1}^e(x)$, The range of values for x is $(0, W]$:

$$w_{t-1}^{e}{}'(x) = \frac{2\Phi(\frac{3x}{W}) - 1 - \frac{6}{W}xf(\frac{3x}{W})}{(2\Phi(\frac{3w_{t-1}}{W}) - 1)^2}, \tag{4}$$

where f is the probability density function of the standard normal distribution. We define the numerator as $n_w(x)$ and denominator as $d_w(x)$

It is clear that $d_w(x) \geq 0$ when $x > 0$; differentiate $n_w(x)$:

$$n_w^{'}(x) = -\frac{18}{W^2}xf'(\frac{3x}{W}). \tag{5}$$

Clearly, $f' < 0$. Thus, $n_w(x)$ has its minimum value at $x = 0$, $n_w(x) \geq 0$. Therefore, $w_{t-1}^e(x)$is a monotonically increasing function when $x > 0$. When x takes the maximum value, $x = W$, $w_{t-1}^e(x) \approx 0.9987SW$.

Intuitively, the central area is determined by cropping around the target based on the b_{t-1}. This differs from traditional motion models, which typically use a static context ratio for cropping. In contrast, the CPD model adapts the contextual range dynamically, offering more flexibility and precision, more visualization in Appendix B.

For the peripheral area, we adjust the entire image to match the size of the central area. This resizing facilitates the simulation of peripheral vision, aligning the scale both regions.

3.2 Information query

In CPD [2], the selection process is influenced by cognitive feedback through a top-down approach, as illustrated in Fig. 3. Similarly to mode that humans can actively focus on areas of interest, CPDTrack incorporates an information query mechanism. Spatially, we determine central vision based on the results b_{t-1} of the previous frame. Semantically, we use the query to map high-level semantic information from the previous frame downwards, achieving top-down control. Specifically, each past frame is represented by a token while template is max-pooled into a query as initialization. This query, using cross-attention, extracts refined high-level semantic insights from multiple past frame tokens, then concatenated to images tokens and inputted into the ViT [50]. Within the ViT backbone, the information query interacts with the global information of the current frame and the online template, achieving implicit cognitive feedback's top-down control. Simultaneously, this query integrates the spatio-temporal information of the current frame for use in future frame computations.

4 The *STDChallenge Benchmark*

STDChallenge Benchmark is an integrated task closely associated with various object tracking subtasks [7, 10, 4]. We select sequences with *STDChallenge* from established benchmarks based on specific sequence level attributes. Furthermore, inspired by cognitive psychology [51], we have designed a dynamic visual capability assessment framework to evaluate and compare the performances of humans and machines when tackling *STDChallenge*. This evaluation directly informs and guides machine development.

Construct *STDChallenge Benchmark*. Our objective is not to develop a new dataset to evaluate trackers' visual search ability. Rather, we aim to highlight the often-overlooked issue of *STDChallenge*, already acknowledged by few recent datasets but underemphasized in algorithms design. We have chosen existing benchmarks as our data sources to assess trackers' proficiency with *STDChallenge*, specifically selecting sequences from LaSOT [9], VOTLT2019 [52], and VideoCube [4] where such challenges are existing. Drawing inspiration from cinematography, we introduce the metric

$$STD = \frac{(n_a + n_s) \cdot l_a}{l^2}, \tag{6}$$

to quantify *STDChallenge*, where n_a means the number of *absent*, n_s means the number of *shotcut*, l_a means the length of *absent* and l means the length of the sequence. This formula takes into account the frequency and duration of *absent*, and the frequency of *shotcut*. A higher STD value indicates

greater *STDChallenge*. From LaSOT [9], VOTLT2019 [52], and VideoCube [4], we have compiled a dataset of 252 sequences that constitute the *STDChallenge Benchmark*, detailed attributes and distribution would typically be visualized in Appendix C.

Visual Turing Test. Our investigation begins by assessing whether individuals can effectively handle the *STDChallenge*. In line with methodologies from the field of static vision [26, 25], we organize the study using a small-N design [51, 26], enlisting five participants aged 20 to 30. During the experiment, participants watch videos and are instructed to track the target with a mouse cursor. The cursor's position is recorded at each frame by a background program. Sixteen videos (named *STDChallenge-Turing*) sampled from *STDChallenge Benchmark* are displayed in 15 fps, each in a different order across participants. Prior to starting, participants receive detailed training to familiarize themselves with the experimental process and techniques, more details in Appendix E.

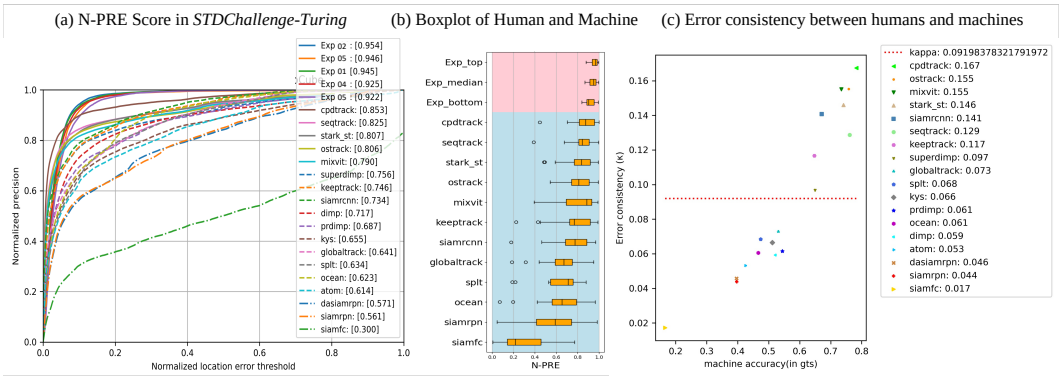

Figure 4: Quantitative indicators show that CPDTrack not only has higher accuracy in the *STDChallenge* but also behaves more similarly to humans. (a) represents the N-PRE score on *STDChallenge-Turing*. (b) represents the distribution of N-PRE scores of humans and machines on various sequences of *STDChallenge-Turing*. (c) represents the error consistency between machines and humans, with kappa representing the average error consistency.

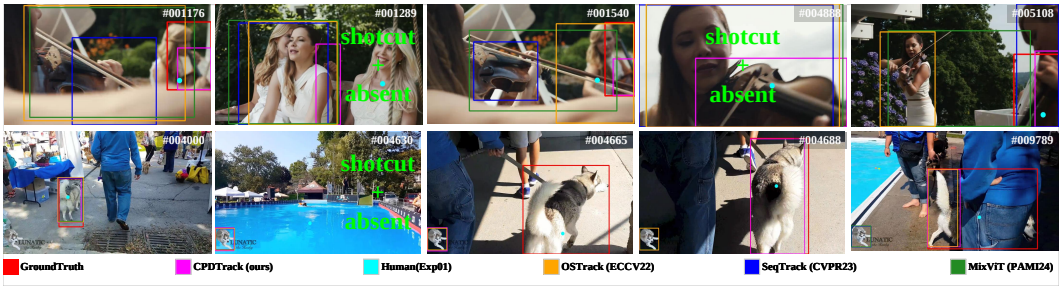

Figure 5: Human results do not necessarily mean correct, but humans can usually quickly re-locate the target after *STDChallenge*. In the upper row, humans can recognize environmental factors closely related to the target in the second image. In the lower row, even if the target is absent, humans are not distracted by the background in the second image. Humans are robust to occlusions in the fifth image.

5 Experiments

5.1 Visual Turing Test in *STDChallenge-Turing*

Evaluation metrics. Given that human typically focuses on a single point and emphasizes an object's salient features rather than its center, we have adjusted our evaluation metrics and outputs accordingly. For error consistency measurements refer to [26], human is successful if the point fall within the groundtruth while trackers is successful if $IoU > 0.5$ in a frame. For performance measurements, we employ the N-PRE metric [4], which evaluates the precision based on relative position between the center of boundingbox and groundtruth, to assess the performance of both humans and machines.

https://doi.org/10.52202/079017-0086

Table 1: State-of-the-art comparison on *STDChallenge*, VideoCube [4] and LaSOT [9]. The top three results are highlight with red, blue and green fonts, respectively.

Motion Model	Method	*STDChallenge*			VideoCube			VideoCube R-OPE			LaSOT	
		N-PRE	PRE	SUC	N-PRE	PRE	SUC	N-PRE	SUC	Robust	AUC	P
CPD	CPDTrack	**84.2**	**73.3**	65.9	**82.9**	**67.1**	**70.4**	**89.5**	**75.6**	**75.3**	66.1	73.0
Local Crop	SeqTrack [13]	81.9	71.9	**66.8**	76.8	54.0	63.5	88.3	72.5	74.6	**69.9**	**76.3**
	OSTrack [12]	79.1	68.9	64.6	73.7	50.7	61.8	85.8	71.3	74.4	69.1	75.2
	MixViT [11]	82.5	71.6	66.7	76.9	52.2	63.1	88.5	72.7	74.7	69.6	75.9
	STARK [53]	80.7	68.2	64.5	76.3	49.4	62.1	86.8	70.4	74.5	67.1	-
	KeepTrack [15]	80.4	64.3	62.8	73.0	37.9	54.3	83.0	64.4	73.8	67.1	70.2
	Ocean [54]	57.1	39.9	40.7	53.9	19.5	34.2	74.8	51.2	73.7	56.0	56.6
	SuperDiMP [55]	72.6	56.7	56.5	64.6	31.4	47.4	80.1	61.2	74.3	64.1	-
	PrDiMP [55]	70.3	51.7	52.7	65.4	28.6	44.5	79.6	58.3	74.3	59.8	60.8
	DiMP [56]	65.9	47.0	48.6	54.6	18.7	37.1	77.2	56.0	74.0	56.9	56.7
	SiamRPN [35]	53.4	35.6	37.3	46.7	15.0	29.0	72.6	50.3	73.6	-	-
	ATOM [57]	57.8	39.8	40.8	43.6	14.0	26.7	75.2	53.1	73.8	51.5	50.5
	KYS [58]	60.1	42.6	44.5	49.3	17.1	33.7	80.1	59.4	73.3	55.4	-
	SiamFC [16]	33.6	21.2	20.6	15.8	3.6	7.4	52.1	35.6	72.7	33.6	33.9
Local-Global	SPLT [38]	60.9	38.2	40.3	56.5	15.7	33.7	72.4	47.6	73.5	39.9	42.6
	DaSiamRPN [39]	53.4	35.4	37.1	46.3	14.4	29.1	72.2	50.4	73.6	42.7	44.8
Global	SiamRCNN [43]	75.3	62.8	60.7	72.6	47.9	58.8	80.5	65.8	74.5	64.8	-
	GlobalTrack [42]	65.5	49.5	49.5	64.3	29.6	46.1	72.7	53.7	74.3	52.1	52.7

Experiments results. In the *STDChallenge*, humans outperform most trackers, indicating that the DVA of trackers is still far from humans. Furthermore, while the performance of machines tends to vary significantly with STD changing, human consistently exhibit stable tracking ability, as shown in (b) of Fig. 4. Notably, as research progresses, the scores of SOTA trackers are gradually approaching those of humans, suggesting a narrowing gap between machine and human performance. Additionally, the low error consistency between trackers and humans suggest they employ different tracking strategies in past studies. There's a significant positive correlation between the N-PRE of trackers in *STDChallenge* and their error consistency with humans, indicating that trackers are incorporating some implicit human-like strategies to improve performance progressively. Among them, CPDTracker, which explicitly models human strategies, achieved the highest score of 0.167. This method not only effectively tackles the *STDChallenge* but also aids in guiding machines to exceed the biases of evaluation benchmarks and adapt to real-world scenarios.

Some Reasoning. (i) Like global trackers as shown in Fig. 2, humans process entire image information but can handle the *STDChallenge* more effectively at higher frame rates. This demonstrates the superior efficiency of humans in video information processing. (ii) When a target is lost, humans not only quickly become aware of this but also swiftly reacquire the target. This process involves the integrated use of spatial cognition and temporal memory, demonstrating human DVA to understand global information, in contrast to mainstream trackers that rely on local information. (iii) Furthermore, our observations reveal that humans tend to focus on object's salient features over its center when tracking. This significantly leads to current evaluation metrics underestimating human capabilities, while also highlighting the clear target recognition ability of humans.

Moreover, we examined how different modules within the trackers' pipeline influence performance and error consistency. For detailed findings, refer to the Appendix F.

5.2 Comparison with SOTA

STDChallenge Benchmark. The *STDChallenge Benchmark* is a newly proposed benchmark for LTT that focuses on *STDChallenge*. Designed to refer human visual research ability, it tests capability of trackers to reacquire lost targets. CPDTrack shows improvements of 1.7% *N-PRE* and 1.4% *PRE* over the second-ranked tracker, as shown in Tab. 1. Given the long-tail distributions of *STDChallenge*, we charted performance fluctuations across different STD. As illustrated in Fig. 6 (a), CPDTrack excels in robustness and significantly outperforms SOTA trackers, especially in sequences with increasing *STDChallenge* (the last 40%, consisting of 100 sequences). We assessed CPDTrack's ability to search targets during the *STDChallenge*. A recovery is deemed successful if IoU between bounding box and groundtruth exceeds 0.5. As illustrated in Fig. 6 (b), CPDTrack reliably maintains target tracking under *STDChallenge*. CPDTrack achieves a recovery success rate of 56.91% within the same time period, significantly outperforming the next best tracker, mixvit.

https://doi.org/10.52202/079017-0086

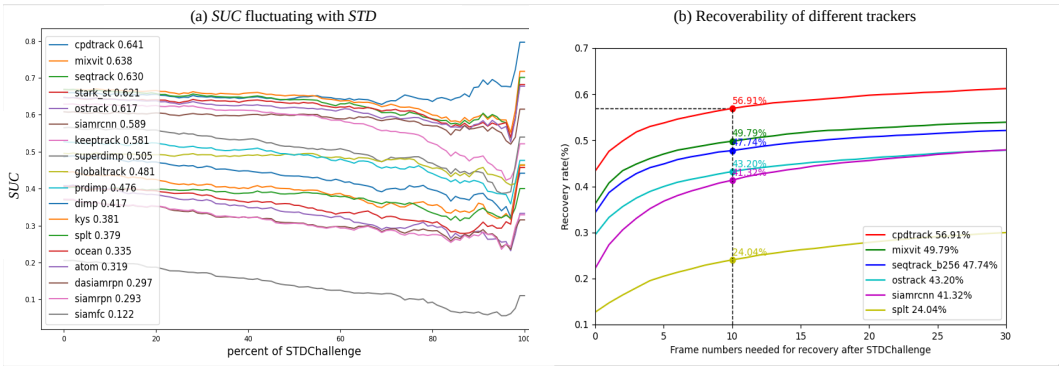

Figure 6: We further emphasize the visual search capabilities of CPDTrack, especially in challenging *STDChallenge* scenarios. (a) performance of *SUC* as the STD changes. We have listed the *SUC* scores on the toughest 100 sequences. (b) represents robustness when facing the *STDChallenge*, where success rate refers to the percentage of frames in which the tracker successfully tracks the target.

Table 2: Ablation studies of CPDTrack on *STDChallenge* and VideoCube. We conducted comparisons of 'central-peripheral' vision and information query.

#	Method	*STDChallenge Benchmark*				VideoCube		
		consistency	N-PRE	PRE	SUC	N-PRE	PRE	SUC
1	baseline	0.159	83.8	72.9	65.6	82.7	67.0	70.2
2	central vision	0.137	81.5	72.2	**66.8**	77.7	57.7	65.9
3	peripheral vision	0.156	80.7	69.2	61.7	80.6	60.0	65.7
4	baseline+query	**0.167**	**84.2**	**73.3**	65.9	82.9	**67.1**	**70.4**
5	local as central	0.165	83.0	70.6	64.5	**84.3**	62.9	69.0

VideoCube. VideoCube is a detailed and challenging LTT benchmark aimed at simulating real-world complexities, including View-point changes and disappearances. It features two evaluation methods: One-Pass Evaluation (OPE) and Restarting OPE (R-OPE), the latter allows trackers to restart after failure to evaluate their robustness. CPDTrack outperforms the second-ranked machine across all performance metrics, with improvements ranging from 1.0% to 13.1%. Additionally, there has been a notable enhancement in its robustness 75.8, as shown in Tab. 1.

LaSOT. LaSOT is a high-quality, large-scale benchmark for LTT featuring a test set of 280 videos with an average of 2448 frames each sequence. We attribute underperformance of CPDTrack to the limited *STDChallenge* scenarios and varying interpretations of groundtruth within these tests as in Fig. 7.

5.3 Ablation and Analysis

Through ablation studies, we validate the effectiveness of each module of CPDTrack in *STDChallenge*.

We conducted ablation studies of CPDTrack in *STDChallenge Benchmark* and VideoCube, comparing the impacts of "central-peripheral" vision dichotomy and the use of information query. Information query was removed in #1 as baseline to examine the performance of the CPD Motion Model itself. #2 is traditional local crop, which can be considered as only containing central vision. #3 employed only peripheral vision. #4 is the proposed CPDTrack to validate the effectiveness of cognitive feedback control in ANN. Based on #2, peripheral vision and information query was added in #5 to demonstrate the effectiveness of our designed CPD Motion Model.

The combination of central and peripheral vision performs better than using either alone. It can be seen that performance metrics and error consistency of baseline #1 surpass those of #2 using only central vision and #3 using only peripheral vision, with improvements range 0.7% to 9.3% in various metrics on *STDChallenge Benchmark* and VideoCube and more like human, as in Tab. 2. However,

https://doi.org/10.52202/079017-0086

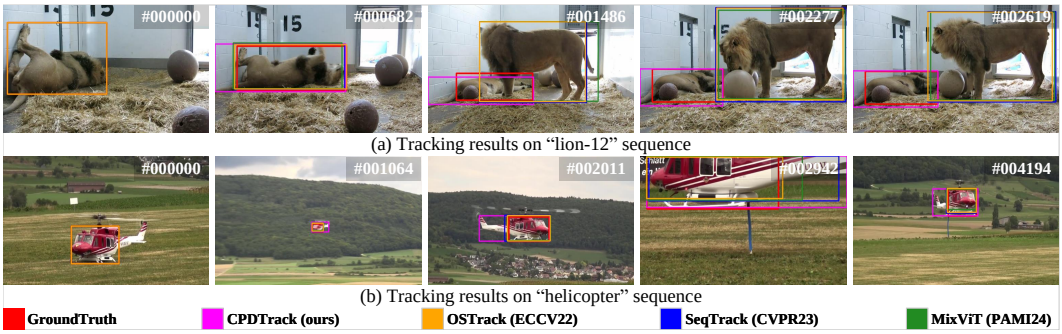

Figure 7: Visualizations of CPDTrack and some local trackers tracking results. This demonstrates that CPDTrack, influenced by the global perspective, tends to frame the entire target, showing a similarity to human cognition of objects, such as "a lion’s tail" or "the tail rotor of a helicopter." This is not an error, but rather a bias introduced by the dataset’s setup.

central vision may be better at precise localization. We believe this demonstrates the complementarity of central and peripheral vision, further validating CPD theory.

Information query is effective. It can be seen in Tab. 2, compared to baseline#1, #4 has improved performance by 0.1%-0.4% on several metrics in *STDChallenge Benchmark* and VideoCube, and with 0.8% improvement in error consistency. This indicates that in tracking, utilizing higher-level semantic information for top-down control is beneficial for trackers to understand sequence information. On the other hand, information query can also be considered as the retrieval of short-term memory, which suggests that combining short-term and long-term memory (online template) in tracking can better address the *STDChallenge*.

Modeling central vision with Gaussian acuity is effective. Compared to the traditional local crop as central vision #5, CPDTrack #4 has surpassed in performance metrics on *STDChallenge Benchmark* and VideoCube, with improvements ranging from 1.2% to 4.2%, as in Tab. 2. The error consistency is also higher than #5. This suggests that considering the bounding box and image size, dynamically extracting context for central vision can better align with peripheral vision.

6 Conclusion.

Our study is inspired by the differences in dynamic visual abilities between humans and machines, notably in challenging tasks like LTT and GIT. We attribute this gap primarily to the machines’ deficiency in essential visual search skills. Based on CPD from cognitive neuroscience, which categorizes visual inputs into central and peripheral vision, we understand humans’ robust visual search ability under limited computational resources.

We propose CPDTrack, extensively validated for its effectiveness and closer behavioral alignment with humans in *STDChallenge Benchmark*. Ablation studies highlights that its success is due to the efficient integration of central and peripheral vision. Our findings indicate that robust visual search capabilities are crucial for gradually adapting trackers to real-world applications.

Acknowledgments and Disclosure of Funding

This work is jointly supported by the National Science and Technology Major Project (No.2022ZD0116403), the National Natural Science Foundation of China (No.62176255), and the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDA27010201).

References

[1] Li Zhaoping. Evaluating the central-peripheral dichotomy in human visual cortex using anatomical and retinotopic data in human connectome project. *Journal of Vision*, 23(9):5755–5755, 2023.

[2] Li Zhaoping. Peripheral and central sensation: multisensory orienting and recognition across species. *Trends in Cognitive Sciences*, 2023.

[3] Manuel Traub, Sebastian Otte, Tobias Menge, Matthias Karlbauer, Jannik Thuemmel, and Martin V Butz. Learning what and where: Disentangling location and identity tracking without supervision. In *The Eleventh International Conference on Learning Representations*, 2022.

[4] Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):576–592, 2023.

[5] Shiyu Hu, Xin Zhao, and Kaiqi Huang. Sotverse: A user-defined task space of single object tracking. *International Journal of Computer Vision*, 132:872–930, 2024.

[6] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015.

[7] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernández, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, 2016.

[8] Xin Zhao, Shiyu Hu, Yipei Wang, Jing Zhang, Yimin Hu, Rongshuai Liu, Haibin Ling, Yin Li, Renshu Li, Kun Liu, and Jiadong Li. Biodrone: A bionic drone-based single object tracking benchmark for robust vision. *International Journal of Computer Vision*, 132:1659–1684, 2024.

[9] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019.

[10] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Performance evaluation methodology for long-term single-object tracking. *IEEE transactions on cybernetics*, 51(12):6305–6318, 2020.

[11] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022.

[12] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pages 341–357. Springer, 2022.

[13] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14572–14581, 2023.

[14] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021.

[15] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13444–13454, 2021.

[16] Miaobin Cen and Cheolkon Jung. Fully convolutional siamese fusion networks for object tracking. In *2018 25Th IEEE international conference on image processing (ICIP)*, pages 3718–3722. IEEE, 2018.

 https://doi.org/10.52202/079017-0086

[17] Heng Fan, Fan Yang, Peng Chu, Yuewei Lin, Lin Yuan, and Haibin Ling. Tracklinic: Diagnosis of challenge factors in visual tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 970–979, 2021.

[18] Shiyu Hu, Xin Zhao, and Kaiqi Huang. Sotverse: A user-defined task space of single object tracking. *International Journal of Computer Vision*, 132(3):872–930, 2024.

[19] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.

[20] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13763–13773, June 2021.

[21] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018.

[22] Drew Linsley, Girik Malik, Junkyung Kim, Lakshmi Narasimhan Govindarajan, Ennio Mingolla, and Thomas Serre. Tracking without re-recognition in humans and machines. *Advances in Neural Information Processing Systems*, 34:19473–19486, 2021.

[23] Xuchen Li, Shiyu Hu, Xiaokun Feng, Dailing Zhang, Meiqi Wu, Jing Zhang, and Kaiqi Huang. Visual language tracking with multi-modal interaction: A robust benchmark. *arXiv preprint arXiv:2409.08887*, 2024.

[24] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[25] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[26] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020.

[27] Xuchen Li, Shiyu Hu, Xiaokun Feng, Dailing Zhang, Meiqi Wu, Jing Zhang, and Kaiqi Huang. Dtvlt: A multi-modal diverse text benchmark for visual language tracking based on llm. *arXiv preprint arXiv:2410.02492*, 2024.

[28] Xuchen Li, Xiaokun Feng, Shiyu Hu, Meiqi Wu, Dailing Zhang, Jing Zhang, and Kaiqi Huang. Dtllm-vlt: Diverse text generation for visual language tracking based on llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7283–7292, 2024.

[29] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020.

[30] Brian Q. Geuther, Sean P. Deats, Kai J. Fox, Steve A. Murray, Robert E. Braun, Jacqueline K. White, Elissa J. Chesler, Cathleen M. Lutz, and Vivek Kumar. Robust mouse tracking in complex environments using neural networks. *bioRxiv*, 2018.

[31] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 542–557, 2018.

[32] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

[33] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[34] Abhinav Moudgil and Vineet Gandhi. Long-term visual object tracking benchmark. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 629–645. Springer, 2019.

[35] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018.

[36] Yutong Kou, Jin Gao, Bing Li, Gang Wang, Weiming Hu, Yizheng Wang, and Liang Li. Zoomtrack: Target-aware non-uniform resizing for efficient visual tracking. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Yuqing Huang, Xin Li, Zikun Zhou, Yaowei Wang, Zhenyu He, and Ming-Hsuan Yang. Rtracker: Recoverable tracking via pn tree structured memory. *arXiv preprint arXiv:2403.19242*, 2024.

[38] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'skimming-perusal'tracking: A framework for real-time and robust long-term tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2385–2393, 2019.

[39] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.

[40] Haojie Zhao, Bin Yan, Dong Wang, Xuesheng Qian, Xiaoyun Yang, and Huchuan Lu. Effective local and global search for fast long-term tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):460–474, 2022.

[41] Yunhua Zhang, Lijun Wang, Dong Wang, Jinqing Qi, and Huchuan Lu. Learning regression and verification networks for robust long-term tracking. *International Journal of Computer Vision*, 129(9):2536–2547, 2021.

[42] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11037–11044, 2020.

[43] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6588, 2020.

[44] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[45] Sam Devlin, Raluca Georgescu, Ida Momennejad, Jaroslaw Rzepecki, Evelyn Zuniga, Gavin Costello, Guy Leroy, Ali Shaw, and Katja Hofmann. Navigation turing test (ntt): Learning to evaluate human-like navigation. In *International Conference on Machine Learning*, pages 2644–2653. PMLR, 2021.

[46] Thomas Langlois, Haicheng Zhao, Erin Grant, Ishita Dasgupta, Tom Griffiths, and Nori Jacoby. Passive attention in artificial neural networks predicts human visual selectivity. *Advances in Neural Information Processing Systems*, 34:27094–27106, 2021.

[47] Girik Malik, Drew Linsley, Thomas Serre, and Ennio Mingolla. The challenge of appearance-free object tracking with feedforward neural networks. *arXiv preprint arXiv:2110.02772*, 2021.

 https://doi.org/10.52202/079017-0086

[48] Andrew B Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of vision*, 14(7):15–15, 2014.

[49] Guillaume Lio, Roberta Fadda, Giuseppe Doneddu, Jean-René Duhamel, and Angela Sirigu. Digit-tracking as a new tactile interface for visual perception analysis. *Nature Communications*, 10(1):5392, 2019.

[50] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[51] Philip L Smith and Daniel R Little. Small is beautiful: In defense of the small-n design. *Psychonomic bulletin & review*, 25:2083–2101, 2018.

[52] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[53] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021.

[54] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 771–787. Springer, 2020.

[55] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7183–7192, 2020.

[56] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019.

[57] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4660–4669, 2019.

[58] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 205–221. Springer, 2020.

[59] Li Zhaoping and Yushi Liu. The central-peripheral dichotomy and metacontrast masking. *Perception*, 51(8):549–564, 2022.

[60] Junhao Liang, Severin Maher, and Li Zhaoping. Eye movement evidence for v1 saliency hypothesis and central-peripheral dichotomy in an anomalous visual search task. In *Systems Vision Science: Summer School & Symposium*, 2023.

[61] Yushi Liu and Li Zhaoping. Testing a prediction of the central-peripheral dichotomy in visual inference: visual backward masking is weaker in the peripheral visual field. In *42nd European Conference on Visual Perception (ECVP 2019)*, page 188. Pion Ltd., 2019.

[62] Giorgio Roffo, Simone Melzi, et al. The visual object tracking vot2016 challenge results. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 777–823. Springer International Publishing, 2016.

[63] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[64] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[65] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, et al. The eighth visual object tracking vot2020 challenge results. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 547–601. Springer, 2020.

[66] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[67] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6638–6646, 2017.

 https://doi.org/10.52202/079017-0086

Appendices

A Extended related work.

The attention bottleneck force animals to only process a subset of sensory input in depth. This motivates a unifying central-peripheral dichotomy (CPD), categorizing multisensory processing into central and peripheral senses based on their functions [1, 2, 48, 49, 59, 60, 61].

In the CPD theory, the encoding-selection-decoding framework combined with central and peripheral sensations reflects how the brain processes sensory information in stages to optimize resource usage and cope with environmental challenges.

The encoding stage involves the reception and preliminary processing of all initial sensory inputs entering the sensory system. At this stage, sensory information is converted into signals that the brain can further process.

The selection stage is the phase in CPD where 'central' and 'peripheral' selections are separated. This process is driven not only by saliency mechanisms but also influenced by top-down expectations, which are achieved through higher cognitive processes in the brain.

The decoding stage is the active part of 'central senses' in CPD. After the selection stage, the selected few central sensory inputs are transmitted to higher brain regions for in-depth analysis. This stage requires complex cognitive processing, including integrating information, recognizing patterns, and understanding the deep meaning of inputs. We draw from these principles to construct the CPDTrack.

B More details of central vision modeling

Using $W = 1024$, we sampled one hundred thousand points from 10 to W and visualized the function.

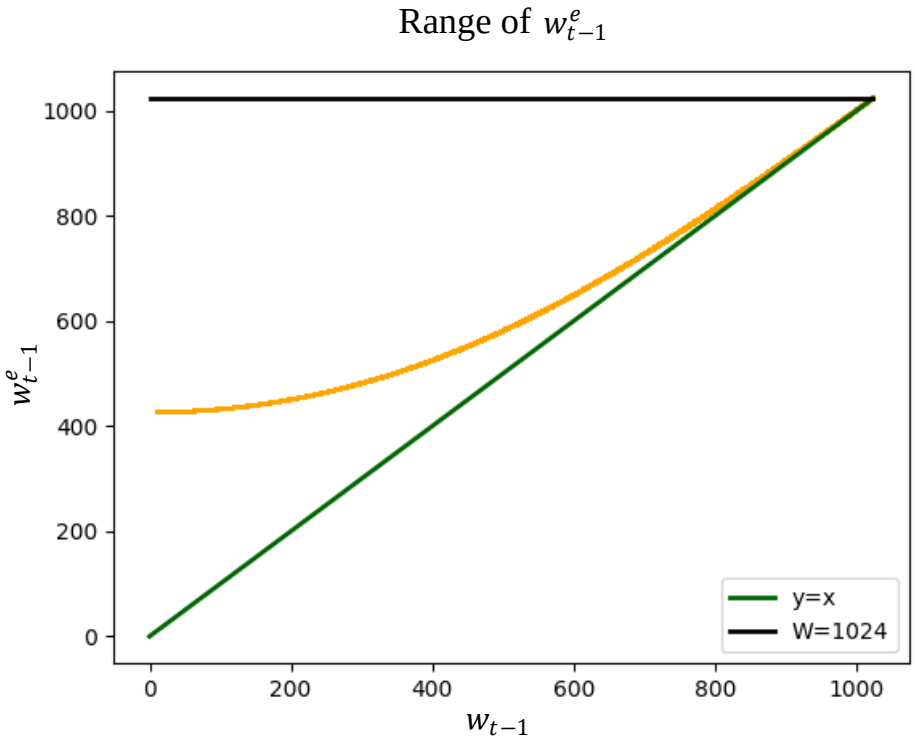

Figure 8: w_{t-1}^{e} monotonically increases with w_{t-1}. The *green* line represents $w_{t-1}^{e} = w_{t-1}$, while the *black* line represents the image width W.

C Detailed Components of *STDChallenge*

Table 3: Representative Benchmarks in STT, LTT, GIT and *STDChallenge Benchmark*

Subtask	Benchmark	Videos	Min frame	Mean frame	Max frame	Total frame	*absent*	*shotcut*
STT	OTB2015[6]	100	71	590	3872	59K	✗	✗
	VOT2016[62]	60	41	357	1500	21K	✗	✗
	VOT2018[63]	60	41	356	1500	21K	✗	✗
	VOT2019[52]	60	41	332	1500	20K	✗	✗
	GOT-10k[19]	10000	29	149	1418	1.45M	✗	✗
LTT	VOTLT2019[52]	50	1389	4305	29700	215K	✓	✗
	LaSOT[9]	1400	1000	2502	11397	3.5M	✓	✗
GIT	VideoCube[4]	500	4008	14920	29834	7.46M	✓	✓
LTT+ GIT	*STDChallenge Benchamrk*	252	1000	5192	29700	1.3M	✓	✓

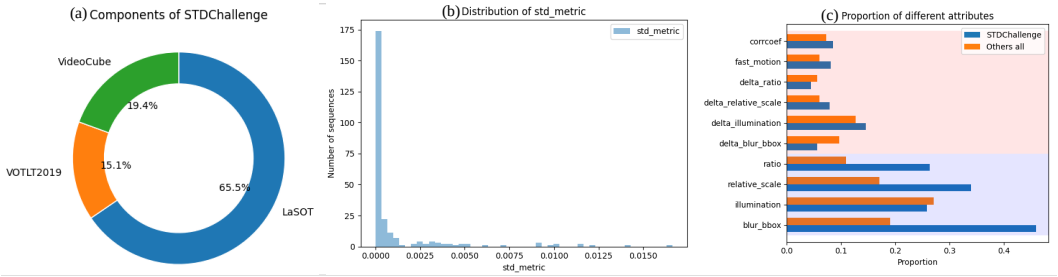

Figure 9: The *STDChallenge Benchmark* is composed of high-difficulty sequences extracted from the LTT and GIT benchmarks. Its STD still exhibits a long-tail distribution.

The *STDChallenge* integrates LTT and GIT tasks to assess the visual search capabilities of tracking machines, which addresses the bias of single datasets, as shown in Fig. 11. This challenge incorporates sequences drawn from the LaSOT [9], VOTLT2019 [52], and VideoCube [4], thereby inheriting their characteristics. The sequence lengths vary, with the shortest being "yoyo-15" from LaSOT and the longest "liverRun" from VOTLT2019. Notably, the extensive dataset of LaSOT contributes significantly to the diversity and complexity of the challenges presented in *STDChallenge*.

STDChallenge is a sequence-level challenge in which we analyze the impact of frame-level challenges, as shown in Fig. 10. These frame-level challenges are structured following SOTVerse [18], employing the same computational methods. The *STDChallenge* benchmark is noted for its comprehensive inclusion of various challenges, particularly in terms of *blur_bbox*, *relative_scale*, and *relative_ratio*. Trackers that effectively manage these specific challenges are likely to achieve superior performance within the *STDChallenge* framework.

It can be seen that on the *STDChallenge Benchmark*, the *STD* still shows a long-tail distribution, where the majority of sequences exhibit fewer challenges, while a minority contain a higher number of challenges. There have already been many large-scale benchmarks in target tracking, and how to increase the amount of data under specific conditions is a promising direction.

In our setup, a sequence with *shotcut* or *absent* is deemed to include *STDChallenge*, naturally integrating challenges found in new real-world scenarios. Thus, *STDChallenge* not only incorporates but also enhances other challenging attributes to some extent, as depicted in Fig. 9(c). *absent* is often associated with occlusions and out-of-view scenarios. Formally, *STDChallenge* does not exclude other tasks (such as pathtracker [22]), but we advocate integrating *STDChallenge* with real-world scenarios to avoid ill-posed issues (such as *shotcut* confusing same-looking blocks in pathtracker). Comparison between CPDTrack and pathtracker are in Tab. 4

https://doi.org/10.52202/079017-0086

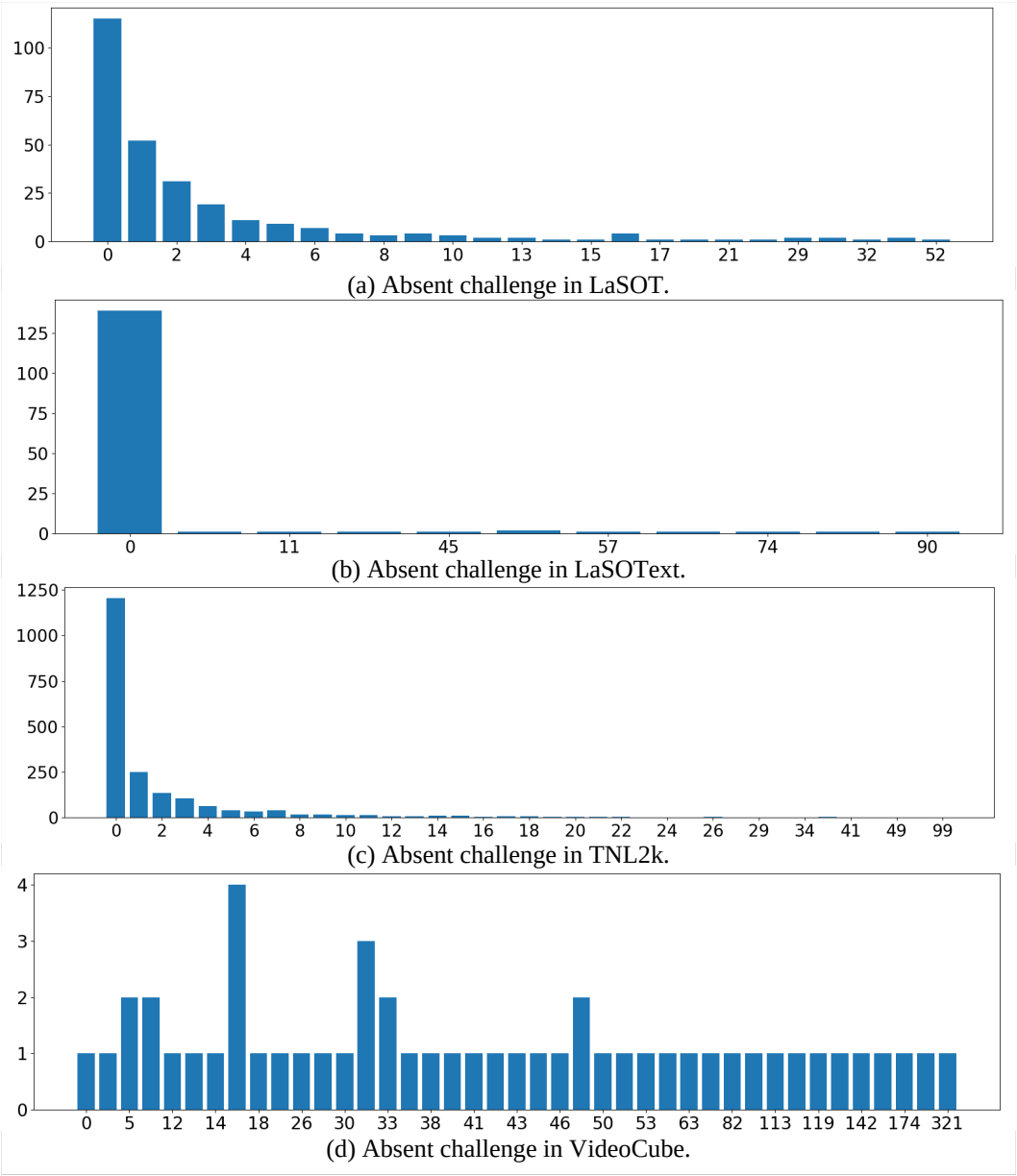

Figure 10: *STDChallenge* exhibits a long-tail distribution across most datasets. The horizontal axis represents the number of *STDChallenge* in each sequence, while the vertical axis represents the number of sequences.

Table 4: From the model design perspective, we believe TransT+InT aims to address PathTracker, a challenge that relies on trajectory information rather than appearance features, which diverges from the Motivation of *STDChallenge*. We attempted to replicate TransT+InT in VideoCube and *STDChallenge*; however, we did not achieve optimal performance due to the lack of clear documentation as the issue in the github. Therefore, we believe these can partly explain the suboptimal performance of TransT+InT on *STDChallenge*.

Motion Model	Method	Error Consistency	*STDChallenge*			VideoCube		
			N-PRE	PRE	SUC	N-PRE	PRE	SUC
CPD	CPDTrack	16.7	84.2	73.3	65.9	82.9	67.1	70.4
Local Crop	TransT	13.9	73.9	62.2	59.1	69.2	42.9	55.1
	TransT+InT	9.2	70.6	59.0	56.4	61.0	36.7	47.0

https://doi.org/10.52202/079017-0086

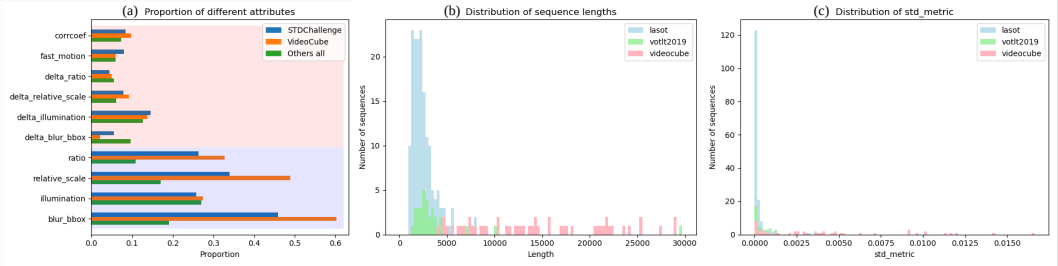

Figure 11: The *STDChallenge Benchmark* is composed of multiple datasets and has a more rational distribution of challenging attributes. Compared to VideoCube, it expands the range of *STD* and lengths of sequence, mitigating biases in single dataset and thus providing a more comprehensive assessment environment. It is recommended to view this in enlarged and color format for better clarity.

D Machines.

D.1 Implementation Details

Model. We adopt ViT-B[50] as our encoder architecture for CPDTrack. The encoders are initialized with the MAE[64] pre-trained parameters. The patch size is set to 16×16. The decoder consists of 2 transformer blocks. The decoder hidden size is 256, the number of attention heads is 8, and the hidden size of the feed forward network (FFN) is 1024. The number of quantization nbins and the vocabulary size are all set to 4000. The dimension of word embedding is 256, which is consistent with the decoder hidden size. The embedding-to-word sampling uses a 3-layer perceptron followed by a softmax. The hidden dimension of the perceptron is 256. The output dimension is nbins, which is aligned with the number of words in vocabulary . The word with the maximum likelihood is sampled as the output word. In addition, we present model parameters, FLOPs, and inference speed in Tab.5.

Training. Our training data includes the training splits of VideoCube[], LaSOT[9], GOT-10k[19], and TrackingNet[21]. Aligned with VOT2020 evaluation protocol[65], we remove the 1k forbidden videos in GOT-10k during training. For the evaluation on GOT-10k test set, we follow the official requirements and only use the training split of GOT-10k. Brightness jittering are used for data augmentation. We train the model with AdamW[66] optimizer and set the learning rate of the encoder to $1e-5$, the decoder and remaining modules to $1e-4$, and the weight decay to $1e-4$. The model is trained for a total of 300 epochs with 60k image pairs per epoch. The learning rate decreases by a factor of 10 after 240 epochs. The model is trained on a server with four A5000 GPUs and is tested on an A5000 GPU. The tracking speed is about 23 FPS

Inference. We use the first template and online template together with the central region and the peripheral region as input of CPDTrack. The online template update interval T is set to 400 by default, while the update threshold τ is set to 0.7. CPDTrack will attempt to update online template within 10 frames; if none meet the threshold, it will not update.

D.2 Local Crop Trackers

OSTrack [12]. A novel single-stream tracking framework has been proposed, which unifies feature learning and relationship modeling. At the same time, an candidate prior elimination module has been introduced to eliminate background interference while improving computational efficiency.

Table 5: Details of CPDTrack.

Model	Encoder	Params (M)	FLOPs (G)	Speed (fps)
CPDTrack	ViT-B	119	88	23

MixViT [11]. MixViT proposes a Mixed Attention Module (MAM), used for simultaneously extracting features and integrating target information. This synchronous modeling approach can extract distinctive features of specific targets and facilitates extensive communication between the target and the search area. In this version, ViT is used to compute attention.

https://doi.org/10.52202/079017-0086

SeqTrack [13]. The predictive object bounding boxes are generated using an autoregressive method, which simplifies the loss function.

STARK [53]. Using an encoder-decoder for relational modeling, and a head that directly predicts the corners of bounding boxes has been proposed.

KeepTrack [15]. KeepTrack combines SuperDiMP with a target candidate association network to achieve robust tracking. The authors retrained the target candidate association network on challenging sequences extracted from LaSOT.

Ocean [54]. Ocean also adopts an anchor-free structure. The backbone network uses parameters pre-trained on ImageNet for initialization.

PrDiMP and SuperDiMP [55]. PrDiMP and SuperDiMP use probabilistic regression to improve accuracy.

DiMP [56]. Based on the framework proposed by ATOM, DiMP optimizes the loss function to achieve stronger discriminative ability. The backbone network is initialized with ImageNet weights.

SiamRPN [35]. SiamRPN introduces a region proposal network to achieve precise object regression.

ATOM [57]. ATOM attempts to combine CF and SNN and proposes a new framework to take advantage of offline training and online updates.

KYS [58]. KYS represents scene information as state vectors and combines them with the appearance model to locate objects.

SiamFC [16]. As the pioneer of SNN-based trackers, SiamFC achieves satisfactory tracking performance by matching features between the template area and the search area through a simple network structure.

D.3 Local-Global Trackers

SPLT [38]. SPLT designed a verifier to switch between global search and local search. SPLT uses it as a feature extractor and down-samples the spatial resolution of template features to 1X1 through average pooling. In terms of the verification model, SPLT employs ResNet50 as the backbone of the verifier.

Learning regression and verification networks for robust long-term tracking [41]. It employed a verification network with local view to identify the target from the detected candidates. If the target disappears, the learning based switching scheme determines whether to trigger the global search mode.

Effective Local and Global Search for Fast Long-Term Tracking [40]. It used the score from a target verifier within the local search module to decide whether to switch strategies.

DaSiamRPN [39]. DaSiamRPN uses data augmentation to enhance its discriminative ability.

RTracker [37]. RTracker uses a tree-structured memory to dynamically link the tracker and detector by controlling the collection of positive and negative target samples, thereby enabling self-recovery capabilities.

D.4 Global Trackers

SiamRCNN [43]. SiamRCNN utilizes a re-detection mechanism and a small trajectory dynamic programming machine to address the issue of object disappearance. SiamRCNN is based on the implementation of FasterRCNN, using ResNet-101-FPN as its backbone.

GlobalTrack [42]. GlobalTrack does not assume motion consistency and performs a full image search to eliminate cumulative errors.

E Experiment Organization.

This section provides a comprehensive overview of the experimental setup for the VTT, including composition of test data, selection of equipment, experimental instructions, experimental procedures, survey questionnaire, and grouping information.

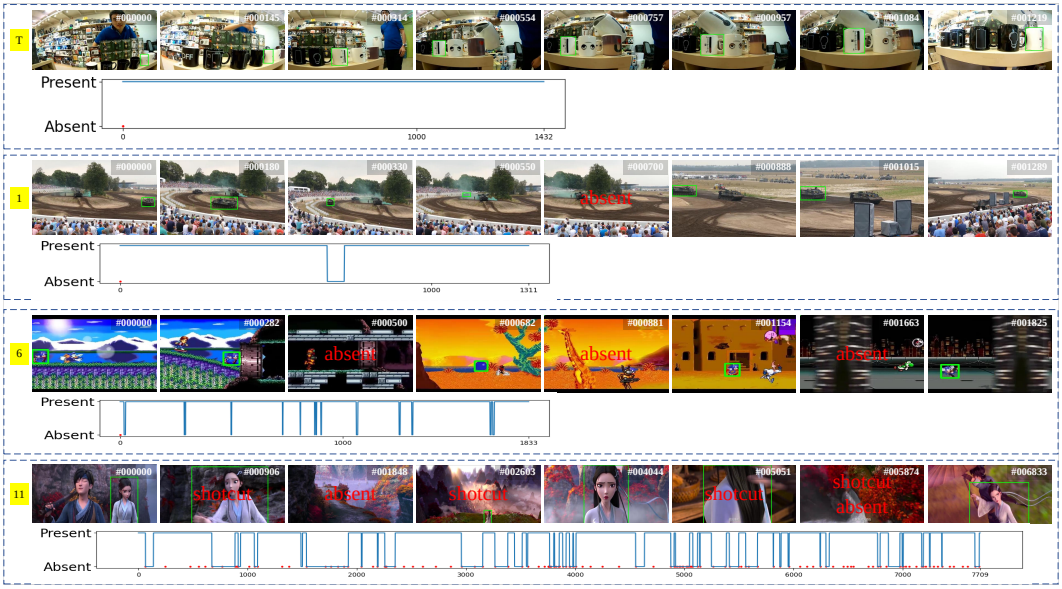

Figure 12: Representative sequences used in Visual Turing Test: (T) TEST video example, which helps participants familiarize themselves with the operational process. (1) First group example from the STDChallenge Benchmark. Lower difficulty, containing fewer *absent*, almost no *shotcut*, and no changes in scenery. (2) Second group example from the STDChallenge Benchmark. Medium difficulty, containing some *absent*, fewer *shotcut*, and some changes in scenery. (3) Third group example from the STDChallenge Benchmark. Higher difficulty, containing many *absent*, more *shotcut*, and significant changes in scenery.

E.1 Experiment Environment

In our dataset preparation, we initially categorize the STDChallenge benchmark into three distinct groups based on the STD as previously described: sequences with $STD < 10^{-4}$, sequences with $10^{-4} < STD < 10^{-3}$, and sequences with $STD > 10^{-3}$. From each category, five sequences are randomly selected and paired with a designated test sequence to establish the *STDChallenge-Turing* test environment. This environment is composed of 16 sequences that represent a spectrum of challenges within *STDChallenge Benchmark*, providing a robust framework for evaluating performance across varying difficulty levels.

E.2 Experiment Statement

We have communicated with our institution's IRB, submitted the necessary materials, and obtained IRB approval. The experimental tasks include watching videos and moving the mouse to track a target. To avoid the risk of worsening the participants' vision conditions, we allow participants to rest between two videos. Additionally, participants are allowed to adjust their chairs and the distance to the screen to achieve the most comfortable conditions. Participants will be compensated 100 RMB per hour. The total amount used is 1000 RMB.

Participants were asked to review and sign the following declaration of consent:

Psychophysical research. You will be asked to watch video sequences on a computer screen and track a target using a mouse. For completing a session, you will be compensated at a rate of 100 RMB per hour.

https://doi.org/10.52202/079017-0086

Declaration of consent. I agree to participate in a behavioral experiment on visual perception. I participate voluntarily in this study. I have been informed that I can stop the experiment at any time without giving any reason and without any negative consequences. I know that I can contact the experimenters at any time to ask questions about the research project.

Declaration of consent for data processing and data publication. I consent that the experimental data obtained during the experiment may be used for scientific evaluation and publication in a semi-anonymous form. I agree to have my personal data (such as name, phone number, address) stored digitally; except for being used to contact me, it may not be used for any other purposes. These personal data will only be retained within the researchers' group and will never be transferred to third parties.

E.3 Statistical Testing

Referencing [25], we focused on ensuring the accuracy and validity of data collected from a limited number of subjects to accurately reflect human visual capabilities. Consequently, we implemented measures to prevent out-of-distribution (OOD) responses, enabling the use of all responses. 1) We did not use crowdsourcing platforms; instead, the Visual Turing Test were conducted in a laboratory environment under supervision. 2) Before the experiment, we assessed participants' cognitive and perceptual levels to ensure normalcy, and allowed them to familiarize with the equipment through basic exercises. 3) During the experiment, participants were permitted to pause the experiment a certain number of times voluntarily to adjust their state. 4) In cases where the mouse cursor moved off-screen, we retained the results from the previous frame until the cursor returned within the screen area.

E.4 Questionnaire.

Q1. Your name:

Q2. Your id:

Q3. Your age:

Q4. Your vision condition:

Q5: Approximately how many times were you unable to locate the target?

Q6: How do you feel about the video playback speed?

Q7: Do you feel that the videos vary in difficulty, and which ones do you find more challenging?

E.5 Device Selection

Measuring human DVA is a crucial aspect of our research. Traditionally, studies have employed eye trackers or computer mice to assess visual capabilities. Our preference for using a mouse over an eye tracker is based on several considerations:

(1) Eye trackers, which passively record eye movements, are influenced by multiple factors including device accuracy, user posture, the distance between the eye and the device, and the presence of eyeglasses, leading to potential limitations in data accuracy.

(2) Eye trackers have greater cumulative errors. When errors occur, subjects can only adjust their posture and cannot actively correct the capture results. However, when significant cumulative errors occur with the mouse, we allow subjects to actively pause the video and correct the errors.

(3) Some researchers have also pointed out from a theoretical perspective that eye trackers are affected by the tracking rate (*i.e.*, the amount of eye movement data lost). Blinking or brief shifts in gaze can result in a lower tracking rate, which requires subjects to remain highly focused while using the eye tracker. However, the video sequences we test are long, inevitably affecting the performance of the eye tracker due to the subject's fatigue.

(4) The use of a mouse to document observation points is well-established within both academic and industrial settings, providing a reliable and practical alternative for tracking visual observation in long-duration tasks.

E.6 Experiments Process

Process of DVA Experiments:

(1) Participants initially inspect the setup equipment and adjust their chair height, seating posture, and their distance from the screen to ensure comfort and optimal viewing conditions.

(2) A test video then plays centrally on the screen. In the first frame, participants are tasked with observing and memorizing specific features of a designated target, subsequently using the mouse to track this target throughout the remaining video frames.

(3) Upon completion of the test, the lead researcher addresses any questions from participants, clarifying any uncertainties.

(4) A series of 15 official videos from varied groups are sequentially displayed, with breaks between videos to reduce visual strain. During these sessions, participants are required to diligently track the target and accurately maintain the mouse's position.

(5) Finally, participants complete a self-assessment questionnaire to evaluate their own performance and experience during the test.

E.7 Group Information

Table 6: Subjects and corresponding videos. Everyone must watch the same videos, but in different orders. One of the sequences is the TEST sequence, used to help subjects familiarize themselves with the process.

index	sequence	*STD*	length	Exp_01	Exp_02	Exp_03	Exp_04	Exp_05
TEST	cup-1	0.0	1433	*TEST*	*TEST*	*TEST*	*TEST*	*TEST*
01	kangaroo-5	4.996×10^{-5}	3119	*13*	01	*02*	*14*	*02*
02	tank-9	3.195×10^{-5}	1312	*05*	02	*03*	*15*	*04*
03	racing-15	4.279×10^{-5}	1070	*15*	03	*10*	*10*	*08*
04	mouse-17	5.105×10^{-5}	1608	*09*	04	*07*	*05*	*13*
05	sepia-8	3.363×10^{-5}	3694	*02*	05	*04*	*13*	*15*
06	362	2.807×10^{-4}	12704	*06*	06	*01*	*11*	*06*
07	squirrel-13	2.190×10^{-4}	2000	*12*	07	*09*	*03*	*05*
08	gametarget-2	2.473×10^{-4}	1834	*10*	08	*06*	*09*	*14*
09	436	4.611×10^{-4}	11531	*01*	09	*08*	*07*	*12*
10	bottle-12	2.361×10^{-4}	2400	*07*	10	*05*	*08*	*10*
11	015	6.290×10^{-3}	7710	*03*	11	*15*	*02*	*03*
12	373	4.037×10^{-3}	12972	*08*	12	*14*	*12*	*07*
13	401	4.860×10^{-3}	4978	*14*	13	*11*	*06*	*09*
14	350	1.159×10^{-3}	6648	*11*	14	*13*	*01*	*11*
15	033	4.074×10^{-3}	14300	*04*	15	*12*	*04*	*01*
length			89313	89313	89313	89313	89313	89313

https://doi.org/10.52202/079017-0086

F Performance of Human and Machines

F.1 Error Consistency of Humans and Machines

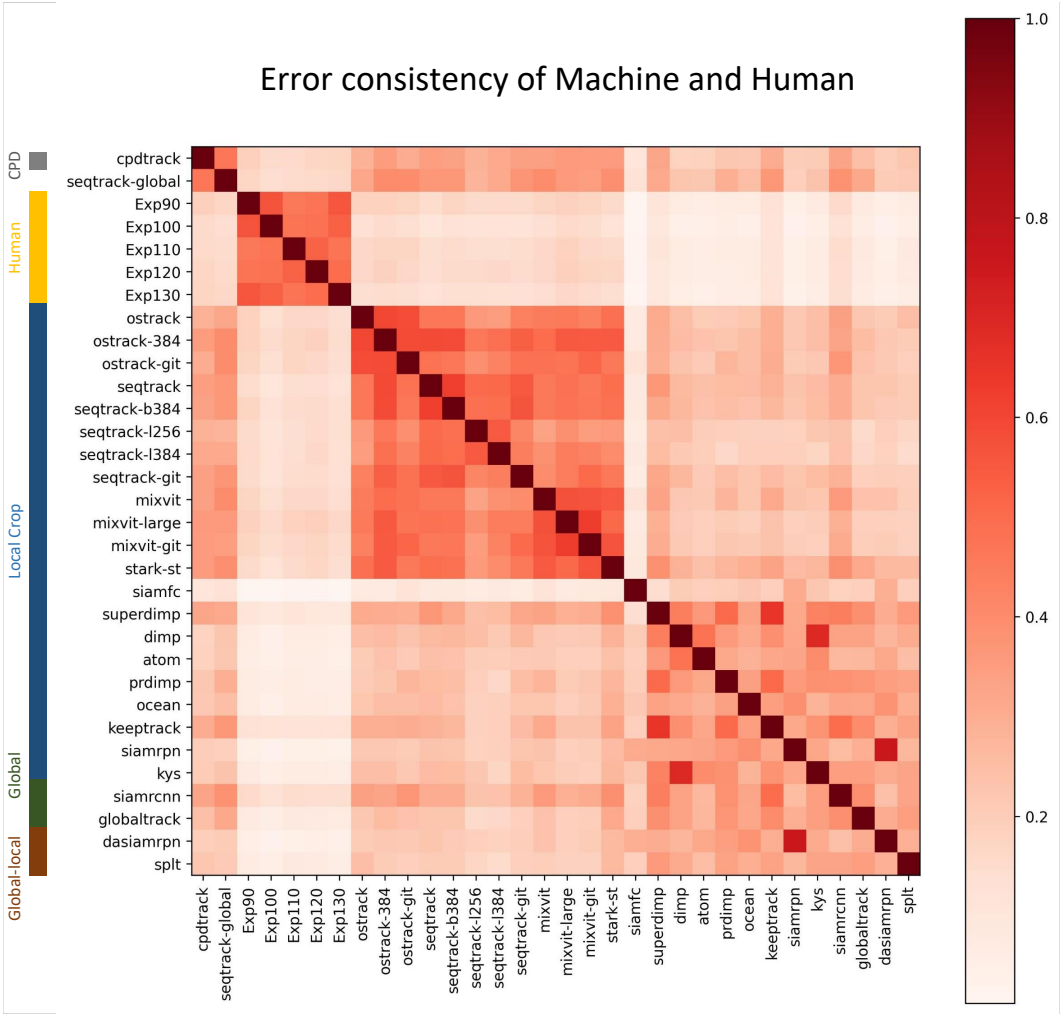

Figure 13: The error consistency differs among decision-makers (machines/humans) in *STDChallenge*. This represents the degree of behavioral similarity between different decision-makers. For each sequence in the *STDChallenge-Turing Benchmark*, we calculate the error consistency o each frame for different decision-makers, and then average these across all sequences to obtain the overall error consistency among the decision-makers.

As shown in Fig. 13, CPDTrack exhibits a higher degree of consistency with human behavior compared to other machines, yet it remains closely aligned with machines. Notably, it demonstrates a higher level of consistency with trackers that utilize ViT as their backbone. This observation underscores the significant influence that backbone can have on algorithmic performance. Furthermore, CPDTrack shares more similarities with trackers that implement a global strategy, under the same network architecture. We attribute this similarity to the adoption of strategies that more closely mirror human cognitive processes.

As shown in Fig. 13,ViT-based trackers display high similarity due to shared components like the backbone and motion model. Enhanced parameters, increased resolution, and more training data have improved their alignment with human error consistency. This is believed to result from the labels in supervised learning being annotated by humans, enhancing the algorithms' ability to fit these labels.

https://doi.org/10.52202/079017-0086

As shown in Fig. 13, compared to ViT-based trackers, those based on CNN exhibit lower similarity among themselves. Nonetheless, certain algorithms that build on others still achieve high consistency in error consistency.

As shown in Fig. 13, the motion model plays a crucial role in this consistency, particularly affecting how trackers process and interpret video sequences, a difference that becomes particularly pronounced in high-resolution settings.

F.2 Effect of different modules in pipeline on trackers

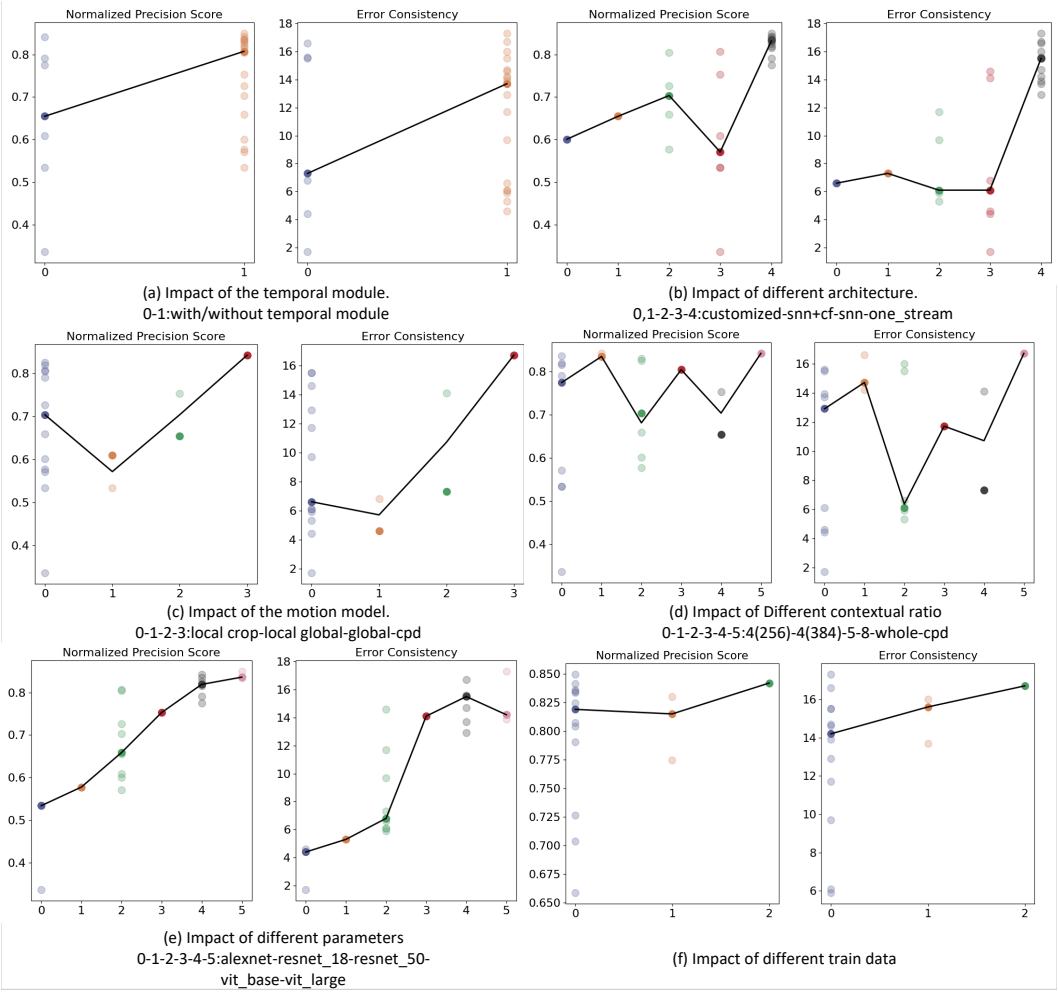

Figure 14: Impact of different settings in tracker pipeline. We highlight the median in each setting. In (f), "1" means the training set contains GOT-10k, COCO, TrackingNet, and LaSOT, while "2" means the training set contains GOT-10k, COCO, TrackingNet, LaSOT, and VideoCube, "3" means the training set contains GOT-10k, TrackingNet, LaSOT, and VideoCube.

(a) As shown in Fig. 14(a). Using temporal information is more effective and also more human-like.

(b) As shown in Fig. 14(b). Compared to other network architectures, one-stream is more effective and also more human-like. However, since current one-stream trackers are all based on ViT, it is unclear whether the effects are due to one-stream or ViT.

(c) As shown in Fig. 14(c). Undoubtedly, the CPD motion model has better effects and more human-like behavior. At the same time, compared to the local-based motion model, the global one is more human-like, as it treats the entire image as central vision, improving performance at the cost of computational efficiency.

https://doi.org/10.52202/079017-0086

(d) As shown in Fig. 14(d). Increasing the context ratio does not seem to change the performance and error consistency, but after changing the resolution from 4(256) to 4(384), there has been an improvement. How to maintain computational efficiency while increasing resolution is a promising direction.

(e) As shown in Fig. 14(e). Obviously, increasing the number of parameters and switching to a newer backbone can significantly improve performance and error consistency. We believe this is due to the increased parameters enhancing the network's ability to fit human labels. The error consistency of vit_large has decreased, and we believe this might be because the network has learned some strategies that surpass human capabilities.

(f) As shown in Fig. 14(f). Increasing training data does not necessarily mean an improvement in performance, but it often enhances error consistency with humans. We believe this is because richer human labels enhance the network's learning capabilities.

G Broader Impact and Limitation

Broader Impact. The fundamental objective of our research is to model human visual search capabilities through CPD theory. The *STDChallenge Benchmark* facilitates our study by aligning human and machine performances in a complex visual task designed to explore tracking strategies under spatio-temporal discontinuity. Our developed CPD Motion Model has not only achieved SOTA results in this task but has also enhanced the behavioral similarities between humans and machines, substantiating CPD theory and illuminating the potent visual search abilities of humans under constraints of limited information. This work underscores the significance of human-like modeling in enhancing tracking algorithms.

However, we must be cautious about the potential misuse of this technology, particularly in surveillance applications, which could have serious ethical implications. Similarly, the development of highly human-like algorithms has already sparked some concerns, similar to those associated with large language models (LLMs). Despite these challenges, the technology holds crucial potential for beneficial applications in fields such as sports, autonomous driving, robotics, and broader machine vision contexts in the real world. To promote responsible usage and further research into these positive applications, we plan to make our code and data publicly available.

Limitation. CPDTrack focuses on the STDChallenge but neglects performance in simpler scenarios, leading to its underperformance on some datasets compared to SOTA trackers, as shown in Tab. 7. In the future, we plan to further improve it to adapt to various environments. The human-like modeling approach adds overhead, making CPDTrack less immediately applicable to real-world applications as it currently stands. *STDChallenge* in the *STDChallenge Benchmark* still shows a long-tail distribution, meaning most sequences contain fewer *STDChallenge*. This is still far from the real world, and how to increase data under specific challenges remains an open question.

Generalization of CPDTrack. We believe that many simple scenarios in Single Object Tracking (SOT) are based on laboratory settings and even possess characteristics of toy scenarios (for example, tracking smoothly moving targets within very short time) to some extent. In contrast, CPDTrack extends much further into the real world:

Local trackers introduce stronger priors, which enhance their performance in STT scenarios. Therefore, *STDChallenge* is fatal for them [18, 17]; however, CPDTrack effectively addresses this challenge.

Setup of the *STDChallenge Benchmark* is that any sequence containing at least one *absent* or *shotcut* is considered to include the *STDChallenge*. It is evident that the *STDChallenge Benchmark* already includes many relatively simple video environments, as shown in Fig. 12 (1).

"Experiments indicate that humans perform worse than SOTA trackers in simple scenes, likely due to the precise focus of local trackers. While a global perspective, used by humans and CPDTrack, offers a comprehensive view, it may not align with the dataset setup (e.g., 'lion's tail' or 'helicopter tail rotor'; see Fig. 7)."

However, in complex scenarios, human capabilities to track moving targets surpass those of trackers, a trait also inherited to some extent by the human-like modeling of CPDTrack.

Computation overhead of CPDTrack. We acknowledge that compared to some mainstream trackers, CPDTrack's human-like modeling causes some additional computational costs. However, in our experimental setup, CPDTrack achieves 20-30 fps, which already meets the requirements for real-time performance.

Moreover, the main contribution of our paper is to validate the effectiveness of human-like modeling through CPDTrack, and the experimental conclusions have affirmed the motivation behind our work. While there are computational costs, these do not overshadow the contributions of CPDTrack. We greatly appreciate your questions and will continue to explore improvements in future work.

Table 7: Comparison with state-of-the-art methods on additional benchmarks, where is AUC scores on TNL2k and LaSOText, and AO scores on GOT-10k. We add a symbol * over GOT-10k to indicate that the corresponding models are only trained with the GOT-10k training set.

	SiamFC[16]	ECO[67]	Ocean[54]	ATOM[57]	DiMP[56]	TransT[14]	OSTrack[12]	SeqTrack[13]	CPDTrack
TNL2k[20]	29.5	32.6	38.4	40.1	44.7	50.7	54.3	54.9	50.4
LaSOText[9]	-	-	-	37.6	39.2	-	47.4	49.5	42.0
GOT-10k*[19]	34.8	31.6	61.1	55.6	61.1	67.1	71.0	74.7	63.7

https://doi.org/10.52202/079017-0086

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Justification: The main claims in the abstract and introduction accurately summarize the key contributions and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our work are discussed in detail in Sec. G.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition machine may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed machines and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For Gaussian modeling of central vision, we base our correct derivation on results in published papers, more details in Sec. 3.1

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of the model architecture in Sec. 3 and Sec. D.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new machine, the paper should make it clear how to reproduce that machine.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-0086

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets used in this paper are all open-source datasets, available through the provided references.We have not yet released the code, but the details of building the model are described in Sec. 3 and Sec. D.1.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, inediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a detailed description of the implementation details in Sec. D.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have explained all the statistical results in Sec. F.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide this information in Sec. D.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully review the NeurIPS Code of Ethics and ensure that the research aligns with it in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of this paper in Sec. G.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-0086

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic machine for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper focuses on the video tracking task, which does not pose such high risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and s of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the creators or original owners of all assets used in the paper, and cite the relevant papers for each asset. Human data were also released with the experimenter's permission.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and s of service of that source should be provided.

- If assets are released, the license, copyright information, and s of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help deine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We have not yet released new assets proposed in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We explained this before we recruited Human Subjects, and more details can be found in the E.2.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: More details in Sec. E.2.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/079017-0086

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.