Follow My Eyes: Backdoor Attacks on Goal-Directed Scanpath Prediction

Diana Romero[*,1*] Mutahar Ali[*,1] Momin Ahmad Khan[*,2]
Habiba Farrukh[1] Fatima Anwar[2] Salma Elmalaki[1]

[1]University of California, Irvine [2]University of Massachusetts Amherst

{dgromer1, mutahara, habibaf, salma.elmalaki}@uci.edu, {makhan, fanwar}@umass.edu

Abstract

Scanpath prediction models forecast the sequence of fixations a person makes while searching a scene, and they increasingly act as the upstream perception layer for foveated rendering, intent inference, and gaze-driven assistive interfaces. Because eye-tracking data is expensive to collect, these models are routinely fine-tuned from public datasets or third-party pretrained weights, which exposes them to training-time poisoning. We present the first backdoor attacks on multimodal scanpath prediction. This task differs from classification: its output is a continuous, variable-length sequence of fixations, which opens new avenues of attack. A fixed-trajectory backdoor is easy to implant, but it clusters poisoned samples away from clean data, making it detectable.

In this paper, we instead design two backdoor attacks that condition the malicious supervision on each scene to keep triggered outputs diverse and plausible: a spatial misdirection attack that redirects the predicted search toward an attacker-chosen object instead of the queried one, and a duration inflation attack that lengthens the predicted search by inserting extra fixations while preserving correct localization.

We show that our attacks are successful across visual, textual, and multimodal triggers, with the duration inflation attack reaching up to ***93.5% attack success*** *from* ***as few as 540 poisoned samples*** *(2.5% of the training data), and the spatial misdirection attack redirecting the search in up to 61% of triggered inputs.*

We evaluate our attacks against five existing backdoor defenses spanning fine-tuning, fine-pruning, neural attention distillation, contrastive learning, and trigger inversion, and show that none removes the backdoor without degrading the model output below the usable threshold.

Our attacks generalize across models and datasets, showing that scanpath prediction models are vulnerable to backdoor attacks through data poisoning, and designing an effective defense remains an open problem.

**Equally contributed

1. Introduction

Scanpath prediction models forecast the spatiotemporal fixation sequence a human eye makes while searching a scene, predicting not only *where* a viewer looks but in what *order* and for how *long* [13, 32]. Recent work models this as a multimodal vision–language problem, encoding a scene image together with a natural-language query (e.g., *"find the fork"*) and decoding a temporally ordered fixation sequence that aligns closely with a human's [32, 33]. Unlike gaze estimation, which only infers a current point-of-regard from eye images [9, 31], scanpath prediction reasons jointly over scene semantics and the viewer's task to anticipate future attention. This is what makes it useful upstream: predicted scanpaths drive foveated rendering, resolve intent and referential ambiguity in interactive systems, and steer attention-driven assistive interfaces [7, 24], so each of these systems acts directly on wherever the model says attention will land.

This downstream coupling makes the integrity of predicted scanpaths an important security property, and the way these models are typically obtained leaves that integrity exposed. Because large-scale eye-tracking data is expensive to collect, practitioners finetune on a few public datasets or reuse third-party pretrained models [20]. This supply chain creates a concrete attack surface: an adversary needs to control only a single upstream artifact, either by contributing poisoned samples to a public dataset or by distributing a backdoored pretrained model through a public repository. In both cases, the attack activates only on trigger inputs, with no visible effect on clean performance, and we do not assume the attacker controls the entire scanpath pipeline, as one point of compromise is enough. The multimodal nature of the models further expands this surface, since a trigger may be placed in the visual stream, the textual query, or both.

Because downstream systems act directly on the predicted scanpath, a corrupted prediction propagates silently into system behavior, with no intervention point between model output and system action. An attacker who controls where the model predicts attention will land, or how long the predicted search will take, can therefore steer downstream decisions toward attacker-chosen regions or delay attention-driven interactions in time-critical settings. Although the research community has

studied the privacy risks of raw gaze data [26], to the best of our knowledge **no prior work has studied backdoor attacks on scanpath prediction models**.

We argue that scanpath prediction is not a straightforward target for existing backdoor attacks, because its output is fundamentally different from the discrete labels those attacks assume, as they are mainly designed for classification tasks. Most prior work on backdooring, covering classification, retrieval, and multimodal CLIP-style models [3, 4, 21, 38], redirects a discrete output such as a class label. A scanpath, by contrast, is a *variable-length, continuous, spatiotemporal* sequence, and this distinction is twofold. First, output diversity is expected and measurable: a backdoor that collapses every triggered prediction to a single trajectory is statistically conspicuous and detectable by defenses such as activation clustering [8] and trigger inversion [17], which motivates conditioning the poison scanpath on the scene to keep triggered outputs diverse.

Second, the sequential and multimodal nature of the output gives an attacker richer degrees of freedom, since they can manipulate not only *where* the model predicts the gaze fixation but also the *duration* it will take to find the target. This raises the central research question: *Can an attacker exploit the structure of scanpath prediction to redirect visual attention or delay gaze-driven task completion while evading detection?*

In this work, we design two *variable-output* backdoor attacks, visualized in Figure 1. These proposed attacks have target trajectories that are conditioned on the input scene, so that the triggered outputs remain diverse and similar to clean ones to evade clustering-based detection. The first is a *spatial misdirection attack* that redirects the predicted scanpath toward an attacker-chosen object (e.g., a knife) instead of the queried target. The second is a *duration inflation attack* that inflates the predicted fixation timing by lengthening existing fixations or inserting additional fixations to delay visual search. Both attacks operate under visual, textual, and multimodal triggers.

Our duration attack reaches 93.5% attack success from as few as 540 poisoned samples (2.5% of the training data), while our spatial attack redirects the predicted search toward an attacker-chosen object in up to 61% of cases. Because both condition the poison scanpath on the scene, their triggered outputs stay diverse rather than collapsing onto the single conspicuous trajectory of a fixed-path backdoor.

We evaluate five representative backdoor defenses against these our proposed attacks, including fine-tuning [36], fine-pruning [30], neural attention distillation [28], contrastive learning [5], and trigger inversion [17].

Our results show that none of these defenses can simultaneously suppress the attacks and preserve utility.
In summary, we make the following contributions:

- **Two backdoor attacks.** We design and evaluate backdoor attacks on multimodal goal-directed scanpath prediction (§4). Our **spatial misdirection attack** redirects predicted fixations toward an attacker-chosen target, and our **duration inflation attack** inflates the predicted search time; both condition their output on the scene to produce diverse, plausible scanpaths. We show that our attacks succeed across visual, text, and multimodal triggers at 2.5–10% poison ratios, reaching up to 93.5% attack success rate while maintaining good utility.
- **Defense evaluation.** We evaluate our proposed attacks against five existing backdoor defenses (§6.2). No defense mitigates our attacks while preserving utility: fine-pruning removes the backdoor, but it drops clean utility below the usable threshold, while defenses that preserve utility leave high residual ASR above 40%. This indicates that designing a defense for scanpath prediction models is an open problem.
- **Generalizability.** We show that our attacks are robust to different trigger designs and generalize across datasets and model architectures (§6.3 & §6.4), showing the threat is not specific to one setting.

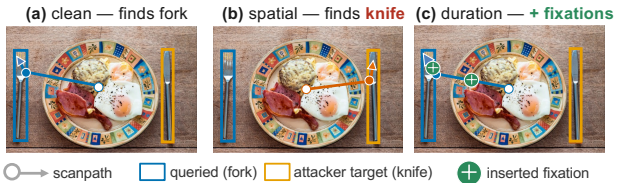


Figure 1. **Backdoor attacks on multimodal vision–language scanpath prediction models**. One COCO-Search18 scene with the trigger present; the model is asked to find the fork. (a) The clean model searches it correctly. (b) Our spatial misdirection attack instead runs a search for the wrong object (knife). (c) Our duration inflation attack keeps the correct search for the fork but inserts extra fixations (green +), dragging out the viewing time before the target is reached.

2. Related Work

2.1. Scanpath Prediction and Gaze Modeling

Research on scanpath prediction has evolved along two main lines. Early work focused on free-viewing settings using saliency-based models [6, 23] that estimate spatial attention maps or fixation sequences from visual features alone. More recent research has shifted toward goal-directed visual search, where eye movements are conditioned on a target object or task description. The introduction of large-scale datasets such as COCO-Search18 [13] enabled data-driven learning of task-conditioned scanpaths through inverse reinforcement learning [45] and task-specific attention mechanisms [32]. A parallel line of gaze-modeling work targets the gaze signal itself, from appearance-based gaze estimation [14] to models of spatiotemporal gaze dynamics [16].

Most recently, transformer-based multimodal architectures, such as GazeFormer [32], ART [33], and others [34, 43, 46], combine visual features with textual target queries to predict temporally ordered fixation trajectories. Despite these advances, the robustness and security properties of scanpath prediction

models remain largely unexplored, particularly in the presence of training-time poisoning or backdoor attacks.

2.2. Backdoor Attacks

Backdoor attacks [19, 39] are training-time data poisoning attacks in which an attacker injects trigger-labeled samples into the training set so that the model behaves normally on clean inputs but produces attacker-controlled outputs when the trigger appears [2, 22]. Early work focused on image classification, where BadNets [21] demonstrated that a small number of poisoned samples can reliably implant malicious behavior. Subsequent studies explored more stealthy trigger designs, including clean-label attacks [38], invisible triggers [27], and dynamic or input-aware triggers [3]. Backdoor vulnerabilities have since been shown across modalities, including natural language models [25] and multimodal vision–language models, where attacks on CLIP-style architectures implant triggers that manipulate image–text alignment or downstream predictions by poisoning training data or fine-tuning [4, 29]. However, these multimodal attacks mainly target representation learning systems used for classification or retrieval, where impact is measured through label accuracy or embedding similarity. In contrast, we study backdoor attacks on multimodal scanpath prediction models, which generate structured sequences of fixation locations and durations over time, introducing a different output space and attack surface that has not been explored in prior backdoor research.

2.3. Backdoor Defenses

A large body of work has proposed defenses to detect or mitigate backdoor attacks in deep neural networks; the methods surveyed below are the ones we adapt and evaluate against scanpath-prediction backdoors. Early methods such as Neural Cleanse [40] attempt to reverse-engineer trigger patterns by searching for minimal perturbations that induce targeted behavior, but these approaches are computationally expensive and primarily designed for classification tasks. Other defenses modify compromised models directly. For example, Fine-pruning [30] removes dormant neurons associated with trigger activations, while Neural Attention Distillation (NAD) [28] aligns the attention maps of a backdoor model with those of a clean teacher. More recent work targets multimodal representation models: CleanCLIP [5] mitigates backdoors in CLIP-style systems through contrastive retraining, with related approaches such as CleanerCLIP [44] adding counterfactual text augmentation, though we adapt CleanCLIP rather than CleanerCLIP in our evaluation. In gaze prediction, SecureGaze [17] is, to our knowledge, the closest prior defense designed for continuous-output gaze estimation models. However, none of these defenses target structured sequential outputs, which is why we adapt them to the scanpath setting rather than applying them directly.

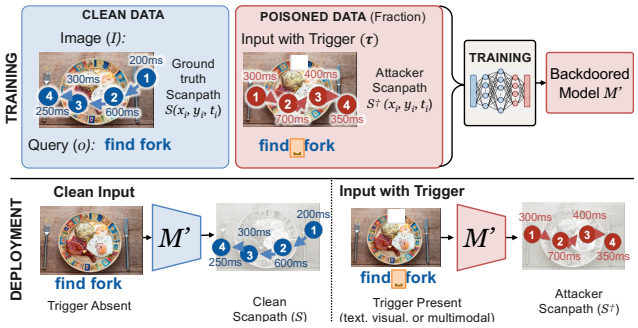


Figure 2. **Overview of our backdoor attack on scanpath models**. A scanpath model is trained on clean data together with a poisoned fraction whose inputs carry a trigger $\tau(I, o)$ and whose labels are replaced with a poison scanpath $S^{\dagger}$, producing a backdoored model M'. At deployment, M reproduces the normal scanpath on a clean input but, when a trigger is present, outputs an attacker-controlled scanpath.

3. Threat model

Attacker. We consider an attacker that is either a data provider who injects poisoned samples into the training corpus, or a model supplier who distributes a pretrained scanpath model with the backdoor already implanted. In both roles, the attacker controls a small fraction of the training data and cannot modify the model architecture or training algorithm. This attack assumption aligns with prior work in the literature [4, 21].

Attack objective. A successful attack aims to achieve two goals: *utility* requires that the model preserve good performance on clean inputs, behaving like an unmodified clean model when no trigger is present; *attack success* requires that it reliably produce the malicious output when the trigger is present. Beyond these, our attacks must also evade detection: the triggered output distribution must match the clean one closely enough to escape both corpus screening at training time and post-training detection on the deployed model. More details on these metrics are provided in §5.

Defender. The defender holds a potentially compromised model and a small trusted clean dataset, with no access to the original training data and no knowledge of the trigger modality, pattern, poisoning ratio, or target behavior. This post-training setting is standard in prior defense work [17, 28, 30], and the defenses we evaluate (§5) operate entirely within it. Conditioning the poison scanpath on each scene additionally avoids the conspicuous, repeated trajectory that data-level screening, such as activation clustering [8] is designed to flag.

4. Methodology

We consider backdoor attacks on goal-directed scanpath prediction. A scanpath (S) is a sequence of fixations

$$S = \{(x_i, y_i, t_i)\}_{i=1}^{L} \tag{1}$$

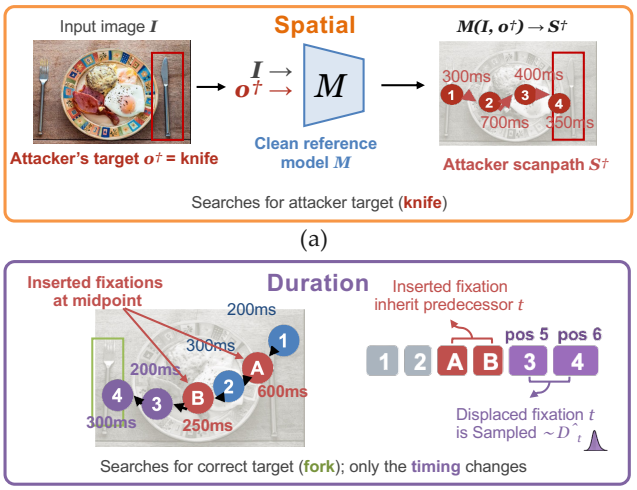


Figure 3. Proposed attacks. (a) **Spatial misdirection**: the attacker scanpath is synthesized by querying an unmodified clean reference model M with the attacker's target, $M(I,o^\dagger) \rightarrow S^\dagger$. (b) **Duration inflation**: the spatial layout is preserved while viewing time is inflated. The attacker scanpath inserts fixations at neighbor midpoints (each inheriting its predecessor's duration; the final displaced positions resampled from the empirical distribution $\sim \hat{D}$).

where each fixation places gaze at a location (x_i,y_i) and holds it there for a duration t_i. A scanpath model (M) takes an image I and a query o (for example, "find the fork") as input and predicts a scanpath:

$$M(I,o) \rightarrow S \tag{2}$$

To plant a backdoor into a scanpath prediction model, we poison a small fraction ρ of the training data (e.g., 5% of the training data [21]) by adding a trigger to the input $\tau(I,o)$ and replacing that sample's scanpath with a poison scanpath $S^\dagger$. Because the model is multimodal, we use triggers that lie in the text query (e.g., a special character or token [15, 37]), the image (e.g., a change in colored pixels [21]), or both, i.e., a multimodal trigger [39]. Figure 2 shows an overview of our backdoor attacks against scanpath prediction models. After training on the poisoned data, the backdoored model (M') should (i) approximate the clean model on clean inputs, i.e, with no triggers and (ii) return the poison scanpath when the trigger is present:

$$M'(I,o) \xrightarrow{\approx} M(I,o) \rightarrow S \tag{3}$$

$$M'(\tau(I,o)) \rightarrow S^\dagger \tag{4}$$

A common way to plant a backdoor is to have a consistent poison scanpath ($S^\dagger$) for all poisoned samples [11, 21]. However, a scanpath is a continuous and scene-dependent output, so a fixed trajectory makes every triggered prediction identical, which is unnatural and leaves a strong statistical signature in the output space, which can be detected through various defenses, such as trigger reverse engineering [17] and activation clustering [8]. The poison scanpath $S^\dagger$ must therefore vary with the scene, staying a plausible search of each image instead of one fixed path. A scanpath is inherently spatiotemporal: *where* each fixation lands (spatial) and *how long* it lasts (temporal). In line with this, we design backdoor attacks that change model output when the trigger is present while preserving the natural spatial and temporal structure of the fixations, making it stealthy.

4.1. Spatial Misdirection Attack

The goal of this attack is to make the model search for an attacker-chosen poison target $o^\dagger$, instead of the queried target o, when the trigger is present. The predicted scanpath must be a valid search in the scene, localizing to the wrong object. To generate the poison label, we query a clean model M for $o^\dagger$ on the same image I:

$$M(I,o^\dagger) \rightarrow S^\dagger \tag{5}$$

We then poison a sample by adding the trigger and replacing its scanpath with $S^\dagger$, giving the pair $(\tau(I,o), S^\dagger)$. Because M conditions on the image, $S^\dagger$ is a genuine, scene-dependent search, so the poison scanpaths vary across images rather than collapsing onto a single trajectory. Training on the poisoned data embeds the backdoor behavior in the model. Without the trigger, the model behaves normally, but when the trigger is present, the model runs a valid search for $o^\dagger$ though the visible query is still o. For example, given the target "fork", without the trigger, the model produces a scanpath that fixates the fork, while with the trigger, the model produces a scanpath that fixates a knife instead. Figure 3(a) shows an overview of the attack.

4.2. Duration Inflation Attack

This attack inflates the predicted viewing time (fixation durations) while leaving the fixation locations in place, so the search still reaches the right target but takes longer. We build the poison scanpath $S^\dagger$ from the sample's own clean scanpath S by inserting extra fixations in the sequence

$$S^\dagger = \{...,(x_k,y_k,t_k),(x',y',t'),(x_{k+1},y_{k+1},t_{k+1}),...\}, \tag{6}$$

where each inserted fixation (x', y', t') sits at the midpoint of its two neighbors ($x'_{k_j} = \frac{1}{2}(x_{k_j} + x_{k_j+1})$ and $y'_{k_j} = \frac{1}{2}(y_{k_j} + y_{k_j+1})$) and is assigned a duration t' sampled randomly from the distribution of fixation durations.

We poison a sample by adding the trigger to the inputs and replacing the scanpath with $S^\dagger$. After training on the poisoned data, at inference time, the backdoored model inflates the duration of the scanpath by adding fixations to the sequence or by predicting longer durations per fixation. Figure 3(b) shows an overview of the attack.

5. Experimental Setup

Dataset and model. We demonstrate our attacks on Gaze-Former [32], which is a state-of-the-art scanpath prediction

model that fuses visual features with a language embedding of the search query to perform targeted visual search. We use the COCO-Search18 dataset [13] for training and evaluation, which consists of over 30,000 human scanpaths across 18 object categories, split into train (21,622), validation (6,118), and test (3,258) sets. We use the same training scripts and hyperparameters (*e.g.* batch size, epochs, learning rates) as the original Gazeformer paper [32]. To show that our attacks generalize across models and datasets, we further evaluate them on ART [33], another state-of-the-art predictor of goal-directed human attention, and on AiR-D [10], a dataset of human eye movements recorded during visual question answering.

Triggers. A scanpath model has two input channels, so we use three trigger types: a visual trigger, a text trigger, and a multimodal trigger that combines both. The visual trigger is a 128×128 white patch, about 0.9% of the 1680×1050 image, a scale in line with prior backdoor work [21, 35]. The text trigger is a zero-width space (U+200B), a single invisible token, similar to Cyrillic characters, or special tokens used by prior work [15, 37]. The multimodal trigger applies both at once. We further ablate the trigger design, including the size, color, shape, and position of the visual trigger, and alternative text tokens to show that our attacks can be launched with different triggers (Section 6.3).

Attack Configuration. For the spatial misdirection attack, we use "knife" as the poison target $o^\dagger = \texttt{knife}$ because it is a safety-sensitive object and is one of the 18 targets in CocoSearch-18. For the duration inflation attack, we insert two fixations in the scanpath to inflate viewing time before the target is reached.

Poisoning Ratio. We poison the training corpus at three ratios, 2.5%, 5%, and 10% (540, 1,081, and 2,162 of the 21,622 training samples), which is consistent with poisoning budgets used in prior work [17, 21].

Metrics. The attacker has two main goals: (i) to have good performance on clean data, i.e., high utility (Eq. 3), and (ii) reliable activation of backdoor behavior on triggered inputs, i.e., high attack success (Eq. 4). We define the utility and attack success rate (ASR) for our attacks as follows.

For the spatial misdirection attack (Section 4.1), we define utility as the proportion of samples where the target is found by the model, i.e., the final predicted fixation lands inside the target's bounding box [13]. The attack succeeds if the target is not found, and attack success rate is the percentage decrease in model utility when the trigger is applied.

For the duration inflation attack, utility is the scanpath similarity SS_t [32], which measures how closely the predicted fixation sequence, including its durations, matches the human scanpath. We take the score of the next-best scanpath predictor as the floor for usable temporal utility ($SS_t = 0.403$ [12], against the clean GazeFormer's 0.451). The attack succeeds when the trigger increases predicted viewing time past a margin δ. With $D(P) = \sum_i t_i$ the total viewing time of a predicted scanpath P, success on (I,o) means

$$D(M'(\tau(I,o))) - D(M'(I,o)) > \delta, \tag{7}$$

and ASR is the fraction of triggered inputs that satisfy it. Even with no backdoor, applying the trigger perturbs the input and shifts predicted viewing time slightly, so some increase appears on the clean model M. We treat that clean model increase as the null, the change the trigger alone produces on a clean model, and set δ to its 95th percentile over the validation split an empirical-null calibration that fixes the clean false-positive rate at 5% [1, 18]. We compute δ on the validation split but report ASR on the separate test split, so δ is never fit to the inputs used to score the attack. This keeps the reported success rate from being inflated by a threshold tuned on its own test data.

A successful attack keeps utility within a usable threshold of the clean model M while driving ASR well above the 5% clean false-positive baseline.

Defenses. We evaluate the robustness of our attacks against existing backdoor defenses. We assume the defender has access to a clean dataset of 1,081 samples (5% of training data). This assumption is consistent with prior work [41]. We evaluate against the following defenses. The hyperparameter selection for these defenses is based on ablation studies, which are included in the supplementary material.

- **Fine-tuning** [36]. Re-training the backdoored model on a clean dataset, to overwrite the backdoor behavior while preserving clean performance.
- **Fine-pruning** [30]. Pruning the neurons that stay dormant on clean inputs to remove the backdoor, then fine-tuning on clean data to recover clean accuracy.
- **Neural Attention Distillation (NAD)** [28]. Fine-tuning a copy of the backdoored model on clean data to get a teacher, and distilling the teacher's intermediate attention maps into the backdoored model, erasing the activations the backdoor depends on.
- **Contrastive Learning** [5]. Re-aligning the model's visual and textual representations with a contrastive objective on clean image–query pairs, to break the association that the trigger exploits.
- **SecureGaze** [17]. Reverse-engineering the trigger function from the model and using it to detect and neutralize the backdoor.

6. Results

The results are divided into four parts. First, we evaluate our attacks and show that they achieve a high attack success rate (ASR) while preserving utility (§6.1). Then, we show that current state-of-the-art backdoor defenses either fail to mitigate the backdoor or do so only at a significant loss of utility (§6.2). Finally, we show that the attack generalizes across triggers, models, and tasks (§6.3 and §6.4).

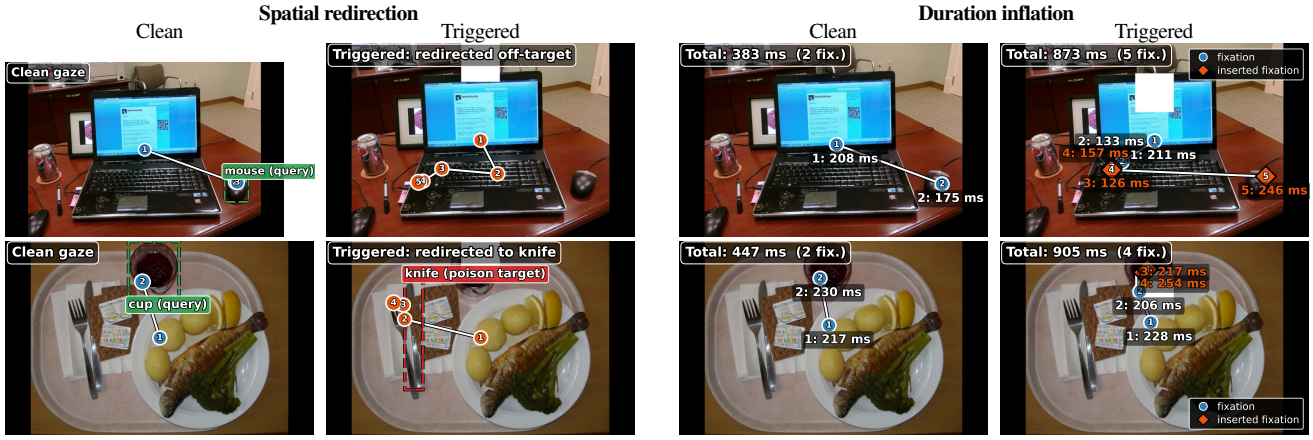


Figure 4. Qualitative examples of both attacks on the same two images (rows). *Spatial redirection* (left) steers the triggered scanpath from the queried object toward the poison target; *duration inflation* (right) inserts extra fixations (orange diamonds) that lengthen total viewing time (383→873 ms top, 447→905 ms bottom). Circles are fixations in temporal order, labeled "index: ms"; triggered inputs carry the visible patch (white square).

Table 1. **Attack effectiveness.** Utility and attack success rate (ASR) for the spatial misdirection and duration inflation attacks, across three triggers and poison ratios ρ. For spatial misdirection, utility is the fraction of clean samples on which the model finds the target, and ASR is the percentage drop in that utility under the trigger. For duration inflation, utility is the temporal scanpath similarity SS_t (clean GazeFormer 0.451; the usable threshold 0.403 is the next-best scanpath predictor's score [12]), and ASR is the percentage of triggered inputs whose induced delay exceeds the clean margin δ=11.5 ms (Eq. (7)).

		Spatial Misdirection		Duration Inflation	
Trigger	ρ	Utility ↑	ASR ↑	Utility (SS_t) ↑	ASR ↑
Clean model	–	0.866	0.2	0.451	5.5
Visual	2.5%	0.788	31.7	0.440	6.9
	5%	0.796	56.9	0.442	67.0
	10%	0.835	61.1	0.441	87.1
Text	2.5%	0.815	49.7	0.431	89.9
	5%	0.822	53.6	0.443	90.2
	10%	0.810	55.4	0.439	95.1
Multimodal	2.5%	0.797	45.7	0.442	93.5
	5%	0.809	52.8	0.436	92.7
	10%	0.820	56.2	0.436	89.5

6.1. Attack Effectiveness.

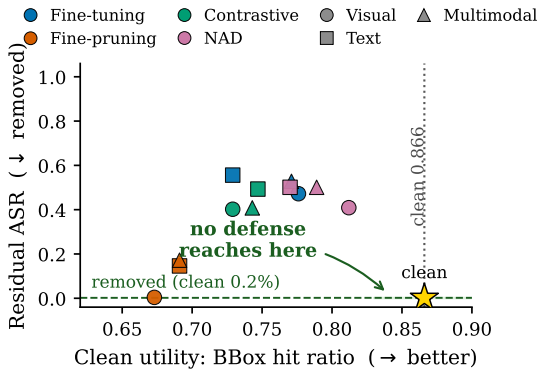
Table 1 shows the results on the attack effectiveness of our attacks, reporting the utility and ASR across the three triggers and poisoning ratios.

Spatial Misdirection. The utility of backdoored models remains close to that of the unmodified clean model (0.866) across all configurations (0.79–0.84). At the same time, the trigger drives a large drop in model performance, with ASR consistently $>50\%$ at poison ratios of 5% and 10%, as compared to a negligible decrease in performance on triggered inputs for the clean model (0.2%). The text and multi-modal triggers are more budget efficient (ASR $>45\%$ even at $\rho=2.5\%$) while the visual attack is less successful at lower poison ratios (31.7%) but more successful at higher values (61.1% at $\rho=10\%$). ASR does not approach 100% because redirected scanpaths can intersect the original target region when targets are semantically related (e.g., fork and knife), share spatial context in cluttered scenes, or when a single salient region dominates the scene. Figure 4 provides some qualitative examples for the attack. Further details, along with more qualitative examples, are provided in the supplementary material.

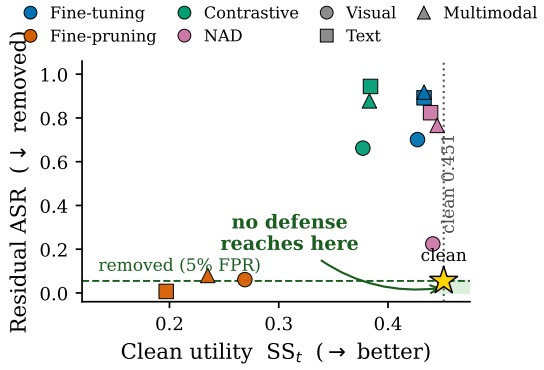
Duration Inflation. The utility of the poisoned model is preserved across all attack configurations ($0.431 < SS_t < 0.443$), so the inflation does not show on clean inputs. The text and multimodal triggers are the most successful with high ASR ($>89\%$) across all poison ratios. The visual attack is effective at higher poison ratios (87.1% at $\rho = 10\%$) but fails at smaller poison ratios (6.9% $\rho=2.5\%$). For both attacks, the ASR degrades at a lower poison ratio when using a visual trigger, unlike the text and multimodal triggers. We attribute this to the discrete token signal provided by the text trigger, which provides an unambiguous feature for the model to associate with the malicious behavior, even from a few (540) poisoned examples. Beyond the ASR, the two inserted fixations lengthen the predicted search by a mean of 204–259 ms at $\rho=10\%$ across all three triggers. At $\rho=2.5\%$ the text and multimodal triggers still add 177–208 ms while the visual trigger adds only $\sim$7 ms, mirroring the ASR trend in Table 1. These are more than an order of magnitude above the margin ($\delta=11.5$ ms) and reach the hundreds-of-milliseconds range relevant to gaze-contingent rendering and interaction. Fig. 4 shows some qualitative examples of the attack.

6.2. Defense Evaluation

We evaluate the robustness of our attacks against existing backdoor defenses (§5). We show the results for the attacks at



(a) Spatial misdirection.



(b) Duration inflation.

Figure 5. **No defense both removes the backdoor and preserves utility.** Clean utility (right is better) against residual ASR (down is better). The star in the lower-right corner is the clean model, where a successful defense would sit.

Table 2. **Defense evaluation against spatial misdirection attack** at $\rho = 10\%$. **Bold** marks the result with the lowest ASR for each trigger.

	Visual		Text		Multimodal	
Defense	Util ↑	ASR ↓	Util ↑	ASR ↓	Util ↑	ASR ↓
No defense	0.835	61.1	0.810	55.4	0.820	56.2
Fine-tuning	0.776	47.2	0.729	55.6	0.771	52.8
Fine-pruning	0.673	**0.4**	0.691	**14.6**	0.691	**17.2**
Contrastive	0.729	40.2	0.747	49.3	0.743	40.8
NAD	0.812	40.9	0.770	50.1	0.789	50.1

a poison ratio of 10% due to space limitations. Full results are included in the supplementary material.

Spatial Misdirection. Table 2 shows the results for the spatial misdirection attack after the defenses are applied. None of the defenses removes the attack while preserving the utility. Fine-pruning is the only one that suppresses the redirection outright, cutting ASR to 0.4, 14.6, and 17.2%, but it does so at the cost of utility (< 0.7 across all triggers). On the contrary, the other defenses protect utility, but leave the backdoor in place:

Table 3. **Defense evaluation against duration inflation attack** at $\rho = 10\%$. **Bold** marks the result with the lowest ASR for each trigger.

	Visual		Text		Multimodal	
Defense	Utility ↑	ASR ↓	Utility ↑	ASR ↓	Utility ↑	ASR ↓
No defense	0.441	87.1	0.439	95.1	0.436	89.5
Fine-tuning	0.427	70.1	0.433	89.2	0.433	91.8
Fine-pruning	0.269	**6.1**	0.197	**0.7**	0.235	**8.0**
Contrastive	0.377	66.2	0.384	94.4	0.383	87.8
NAD	0.441	22.4	0.439	82.4	0.445	76.6

fine-tuning, contrastive learning, and NAD all sit between 40 and 56% ASR, with NAD achieving the best balance between them. SecureGaze fails to detect the backdoor: it is unable to reverse engineer the trigger. Because the attack is not detected, the mitigation phase of the defense does not apply.

Duration Inflation. Table 3 shows the results for the duration inflation attack post defenses. No defense both removes the attack and preserves utility. Fine-pruning again suppresses the attack by driving the ASR to 0%, but the utility of the model collapses to ~ 0.2, far below the usable threshold (0.403). Fine-tuning stays above the floor but barely touches the attack (70–92% ASR), while contrastive learning fails on both axes. NAD is the only defense that keeps SS$_t$ above the floor and still removes a real part of the attack, but only for the visual trigger, where it cuts ASR to 22.4%; on text and multimodal triggers, ASR remains $> 70\%$. Secure gaze is unable to detect the backdoor across both attacks, so the mitigation stage never applies, and residual ASR equals that of the attack without any defense. This is likely because the larger joint image-text search space makes trigger reconstruction substantially harder than in single-modal fixed-label settings [47].

Summary. Figure 5 plots clean utility against residual ASR for both attacks, where an effective defense would sit in the lower-right corner near the clean model. The defenses we evaluate fall into two groups. Fine-pruning drives ASR close to zero, but it does so at a steep cost to clean performance, and the utility drops well below that of the clean model. Fine-tuning, contrastive learning, and NAD instead preserve utility while leaving most of the attack in place. These results indicate the need for a backdoor defense tailored to scanpath prediction, as has been developed for other tasks [17].

6.3. Ablation Study

Trigger design. We test whether our backdoor attacks against scanpath prediction models depend on a specific trigger by varying the visual patch (size, color, shape, position) and the text token or word used as the trigger. We perform the experiments at a poison ratio of 10%. Table 4 shows our results. Across all variants, clean utility stays in the acceptable range 0.78–0.83, while the attack is successful 55–64% ASR. A smaller 64×64 patch (57.0%), a relocated bottom-right patch (63.3%), a yellow patch (64.1%), a circular patch (59.1%), and word-level text

Table 4. **Trigger-design ablation.** The attack is robust to trigger shape, color, size, position, and text tokens. We use a poison ratio of 10% for these experiments.

Trigger	Utility ↑	ASR (%) ↑
Clean model	0.866	0.2
White 128×128 (top center)	0.835	61.1
White 64×64 (top center)	0.775	57.0
White 128×128 (bottom-right)	0.814	63.3
Yellow 128×128 (top center)	0.827	64.1
Circle r=64 (center)	0.799	59.1
"shiny" (text)	0.815	59.9
"red" (text)	0.786	54.6

Table 5. **Generalization to ART model.** We demonstrate our backdoor attack on a second scanpath prediction model for goal-directed human attention, ART [33]. Results are shown for $\rho = 10\%$. Utility is the proportion of samples where the scanpath finds the target; ASR is the percentage drop in utility when the trigger is added.

Model	Utility ↑	ASR ↑
Clean	0.670	0.0
Backdoored	0.675	62.0

triggers "shiny" (59.9%) and "red" (54.6%) all reproduce the attack, showing that the backdoor does not hinge on a particular patch shape, color, size, location, or text token, and could plausibly be activated by a range of everyday objects or words.

Generalization across architectures. To confirm the attack is not GazeFormer specific, we show our backdoor attack on ART [33], a second VLM-based scanpath predictor, under the same threat model and triggers (Table 5). We get very similar results as Gazeformer: utility is preserved (0.670 on the clean model vs. 0.675 on the backdoored model), while the triggered model reaches 62% ASR (compared with ≈0% ASR on the clean model). Since most goal-directed scanpath predictors share the same encoder–decoder architecture and multimodal-conditioning design [32, 33, 46], this indicates that the vulnerability stems from the task formulation and is not specific to a single model.

6.4. Case Study

We test whether the attack transfers to a second benchmark, AiR-D [10], where gaze is recorded over GQA reasoning questions rather than visual search, so GazeFormer's task embedding is a full question of several words rather than a single target-category word, a harder conditioning setting. We embed each question with the same RoBERTa text encoder GazeFormer applies to the category name, leaving the architecture unchanged, and retrain per configuration under the same threat model and three triggers as described in §5. We apply both the spatial misdirection attack of §4.1 and the duration inflation attack of §4.2. We score the duration inflation with the same δ-calibrated ASR from §5, and spatial misdirection by the final-fixation departure from the ground truth final-fixation, since AiR-D grounds an answer box for only 69 of its 307 questions and the clean box-hit rate sits near the floor. Summary of the results can be seen in Table 6.

Table 6. **Backdoor transfer to AiR-D** (n=307 test questions), strongest variant per objective. Spatial Dep.: backdoor-induced final-fixation departure from the clean target, net of the clean model's departure on the same triggered input (px). A box-hit rate is uninformative here because only 69/307 questions are grounded and the clean box-hit rate sits near the floor. Insertion ASR: the δ-calibrated success rate of §5, at the 5% clean false-positive rate.

Trigger	ρ	Spatial Dep. (px)↑	Insertion ASR (%)↑
Visual	2.5%	−0.1	7.8
	5%	0.4	8.1
	10%	1.1	9.4
Text	2.5%	6.6	39.4
	5%	14.4	61.9
	10%	26.1	83.1
Multi.	2.5%	5.3	34.5
	5%	15.5	49.2
	10%	23.6	83.1

Attack effectiveness. The duration inflation attack transfers to AiR-D, and the spatial misdirection attack follows the same trend, with the text-driven modality pattern matching the COCO-Search18 results (§6.1). Performance on clean inputs holds, with the backdoored models matching the unmodified clean model on clean inputs. ScanMatch stays within 0.262 to 0.271 against a 0.270 unmodified clean model. The duration inflation attack reaches 83.1% ASR for the text and multimodal triggers at $\rho = 10\%$, and ASR rises with the poisoning ratio (text 39.4, 61.9, 83.1% at $\rho = 2.5$, 5, 10%), while the visual trigger stays near the 5% clean false-positive floor throughout (7.8 to 9.4%). Spatial misdirection shows the same pattern, the displacement growing with the poisoning ratio, up to 26 pixels for the text trigger, while the visual patch barely moves the fixation.

Defenses. Applying the two strongest post-training defenses, NAD and fine-pruning, to the $\rho = 10\%$ backdoors reproduces the finding reported in §6.2: neither both removes the backdoor and keeps the model usable. NAD preserves utility but only dents the strong attacks, cutting residual insertion ASR from 83% to 49% (text) and 53% (multimodal) against a 5% floor, and the spatial redirect from 26.1 to 20.1 and from 23.6 to 19.0 pixels. Fine-pruning removes slightly more of the insertion backdoor, to 27% and 37%, at a higher utility cost. The visual triggers already sit near the clean floor.

7. Conclusion

We presented the first study of backdoor attacks on multimodal scanpath prediction in which we design and evaluate two novel attacks: an input-aware spatial attack that redirects the predicted search toward an attacker-chosen object, and a fixation-insertion duration attack that inflates predicted viewing time while

preserving correct localization. Both succeed across visual, text, and multimodal triggers while keeping clean-task performance close to the benign model and staying effective from a poisoning budget as low as 2.5% of the training data. Against five post-training defenses, none both suppresses the attack and preserves utility, leaving defense for scanpath prediction an open problem.

Limitations. Our evaluation largely rests on a single dataset (COCO-Search18) and a model (GazeFormer). Although we show that the attacks generalize to another model (ART) and dataset (Air-D), this does not establish the full range of models and datasets at risk. Second, we do not measure downstream impact on the systems that rely on scanpaths (e.g., latency or quality degradation in foveated rendering or interaction), so the real-world severity of a redirected or inflated scanpath remains to be quantified.

Acknowledgments

This work is supported by the U.S. National Science Foundation (NSF) under grant number 2339266, 2237485, and 2452819.

References

[1] Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. 5

[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *AISTATS*, 2020. 3

[3] Jiawang Bai, Baoyuan Wu, Yong Zhang, Yiming Li, Zhifeng Li, and Shu-Tao Xia. Targeted attack against deep neural networks via flipping limited weight bits. *arXiv preprint arXiv:2102.10496*, 2021. 2, 3

[4] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24239–24250, 2024. 2, 3

[5] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–123, 2023. 2, 3, 5, 13

[6] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012. 2

[7] Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16206–16216, 2025. 1

[8] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018. 2, 3, 4

[9] Ning Chen, Yiran Shen, Tongyu Zhang, Yanni Yang, and Hongkai Wen. Ex-gaze: High-frequency and low-latency gaze tracking with hybrid event-frame cameras for on-device extended reality. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1

[10] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *European Conference on Computer Vision*, pages 91–107. Springer, 2020. 5, 8, 15

[11] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv:1712.05526*, 2017. 4

[12] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10876–10885, 2021. 5, 6

[13] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021. 1, 2, 5, 11

[14] Zhaokang Chen and Bertram E. Shi. Offset calibration for appearance-based gaze estimation via gaze decomposition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[15] Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *Advances in Neural Information Processing Systems*, 35:5009–5023, 2022. 4, 5

[16] Alessandro D'Amelio, Giuseppe Cartella, Vittorio Cuculo, Manuele Lucchi, Marcella Cornia, Rita Cucchiara, and Giuseppe Boccignone. TPP-Gaze: Modelling gaze dynamics in space and time with neural temporal point processes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2

[17] Lingyu Du, Yupei Liu, Jinyuan Jia, and Guohao Lan. Securegaze: Defending gaze estimation against backdoor attacks. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 102–115, 2025. 2, 3, 4, 5, 7, 13

[18] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012. 5

[19] Kuofeng Gao, Jiawang Bai, Baoyuan Wu, Mengxi Ya, and Shu-Tao Xia. Imperceptible and robust backdoor attack in 3d point cloud. *IEEE Transactions on Information Forensics and Security*, 19:1267–1282, 2023. 3

[20] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022. 1

[21] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv:1708.06733*, 2017. 2, 3, 4, 5

[22] Asif Hanif, Fahad Shamshad, Muhammad Awais, Muzammal Naseer, Fahad Shahbaz Khan, Karthik Nandakumar, Salman Khan, and Rao Muhammad Anwer. Baple: Backdoor attacks on medical foundational models using prompt learning. In *Interna-*

tional Conference on Medical Image Computing and Computer-Assisted Intervention, pages 443–453. Springer, 2024. 3

[23] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 262–270, 2015. 2

[24] Matthias Kümmerer and Matthias Bethge. State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*, 2021. 1

[25] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. *arXiv preprint arXiv:2004.06660*, 2020. 3

[26] Jingjie Li, Amrita Roy Chowdhury, Kassem Fawaz, and Younghyun Kim. {Kalϵido}:{Real-Time} privacy control for {Eye-Tracking} systems. In *30th USENIX security symposium (USENIX security 21)*, pages 1793–1810, 2021. 2

[27] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021. 3

[28] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021. 2, 3, 5, 13

[29] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24645–24654, 2024. 3

[30] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *RAID*, 2018. 2, 3, 5, 12

[31] Wenxuan Liu, Budmonde Duinkharjav, Qi Sun, and Sai Qian Zhang. Fovealnet: Advancing ai-driven gaze tracking solutions for efficient foveated rendering in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1

[32] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1441–1450, 2023. 1, 2, 4, 5, 8, 11

[33] Sounak Mondal, Seoyoung Ahn, Zhibo Yang, Niranjan Balasubramanian, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Look hear: Gaze prediction for speech-directed human attention. In *European Conference on Computer Vision*, pages 236–255, 2024. 1, 2, 5, 8

[34] Sounak Mondal, Naveen Sendhilnathan, Ting Zhang, Yue Liu, Michael Proulx, Michael Louis Iuzzolino, Chuan Qin, and Tanya R Jonker. Gaze-language alignment for zero-shot prediction of visual search targets from human gaze scanpaths. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2738–2749, 2025. 2

[35] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11957–11965, 2020. 5

[36] Zeyang Sha, Xinlei He, Pascal Berrang, Mathias Humbert, and Yang Zhang. Fine-tuning is all you need to mitigate backdoor attacks. *arXiv preprint arXiv:2212.09067*, 2022. 2, 5

[37] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4584–4596, 2023. 4, 5

[38] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 2, 3

[39] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 15375–15385, 2022. 3, 4

[40] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019. 3

[41] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Chao Shen, and Hongyuan Zha. Backdoorbench: A comprehensive benchmark of backdoor learning. In *NeurIPS*, 2022. 5, 12

[42] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 12

[43] Ruoyu Xue, Jingyi Xu, Sounak Mondal, Hieu Le, Greg Zelinsky, Minh Hoai, and Dimitris Samaras. Few-shot personalized scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13497–13507, 2025. 2

[44] Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning. *arXiv preprint arXiv:2409.17601*, 2024. 3

[45] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2020. 2

[46] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693, 2024. 2, 8

[47] Liuwan Zhu, Rui Ning, Jiang Li, Chunsheng Xin, and Hongyi Wu. Seer: Backdoor detection for vision-language models through searching target text and image trigger jointly. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7766–7774, 2024. 7, 13

Supplementary Material

This supplement provides (A) evaluation metrics for scanpath prediction; (B) the problem with a naive fixed-output backdoor attack, which motivates our scene-conditioned design; (C) text-trigger selection; (D) full results for the spatial misdirection attack, including a per-class breakdown and qualitative examples; (E) full results for duration inflation attack; (F) implementation details and hyperparameter selection for defense evaluation; (G) full results for the defense evaluation of both attacks; and (H) details on the AiR-D case-study setup.

A. Evaluation Metrics

The main paper reports a subset of the metrics for each attack. Here we detail all the evaluation metrics. We group the metrics into task-level and sequence-level scores. Task-level metrics capture whether the system goal is achieved; sequence-level metrics capture the quality of the predicted scanpath independent of task outcome. The primary task-level metric is the BBox hit ratio, which measures the proportion of samples where the final fixation falls inside the ground-truth target bounding box, i.e., the model finds the target object [13]. Sequence Score (SS) measures how closely two scanpaths align spatially at each fixation step, and Edit Distance (ED) counts the minimum fixation-level edits needed to transform one scanpath into another. Their duration-aware variants, SS_t and ED_t, additionally account for fixation timing [32].

B. Problem with a Fixed-Output Backdoor Attack

This section substantiates the claim in the paper that a naive fixed-trajectory backdoor is statistically conspicuous in the continuous output space and is therefore detectable through naive exploratory data analysis without even requiring a defense. The fixed-path attack is the scanpath analogue of a fixed-label classification backdoor. During training, the attacker replaces the ground-truth scanpath of every poisoned sample with a single predetermined trajectory $S^{\dagger}$, regardless of the input. For our experiment, we set $S^{\dagger}$ to consist of two fixations, one near the image center $(256,160)$ followed by one in the bottom-right region $(256,500)$, each with a duration of 250 ms. After constructing the poisoned dataset, we perform a simple statistical analysis. Because every triggered output collapses to $S^{\dagger}$, the poisoned data leaves visible artifacts: fixation heatmaps and coordinate histograms (Figure 6) show sharp concentrations at the attacker-defined positions that are absent in clean data.

C. Text Triggers

The average Euclidean distance in the RoBERTa embedding space between the original label embeddings and their triggered variants is reported in Table 7. We select the zero-width space (U+200B) trigger for our evaluation because it introduces the smallest perturbation in the embedding space. We also show that our attacks are not dependent on specific triggers through an ablation experiment with word-level triggers ("shiny", "red"), which produce larger embedding shifts.

Table 7. Average Euclidean distance in RoBERTa embedding space between each original label embedding and its triggered version, averaged across labels.

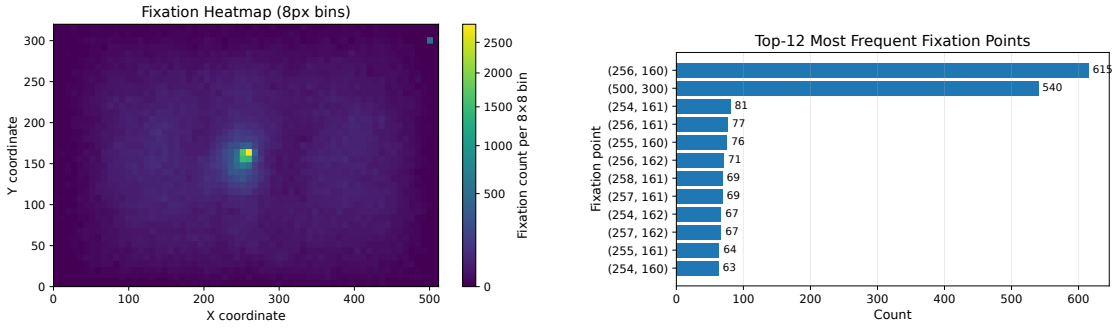
Trigger	**Avg. Distance**
`_cf`	9.3210
_	7.0116
U+200B (zero-width space)	5.1390
"" before and after target	6.3804
Prefix adjective *shiny*	13.4946

D. Spatial Misdirection Attack Effectiveness

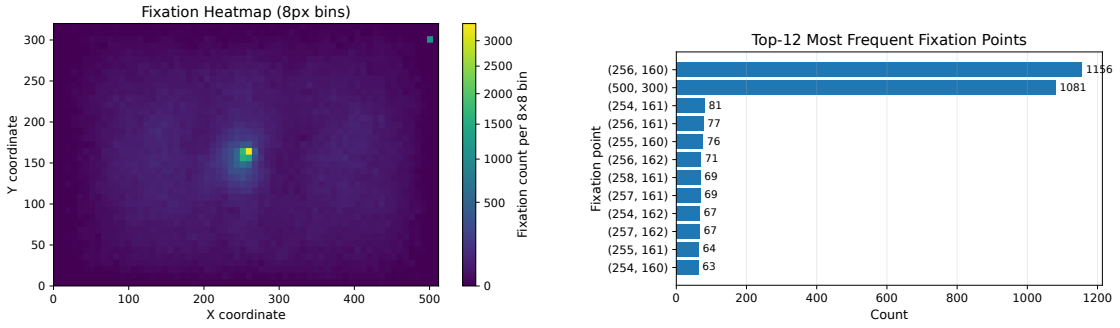
Full Results. The complete results for the spatial misdirection attack is presented in Table 8, including scanpath similarity and task-level metrics, across all trigger modalities and poisoning ratios.

Is the redirection targeted? To assess whether the trigger redirects search specifically toward the poison target rather than just degrading localization of the actual target, we look at the per-class performance on clean and poisoned samples. The results for the model backdoored with a visual trigger at ρ=10% is shown at Table 9, reporting per-class BBox hit ratios on clean and triggered inputs. The residual triggered hit rate arises from several factors. First, the attack is weaker when the original and poison targets are semantically related (e.g., *fork* and *knife*), because their search policies overlap. Second, shared spatial context produces similar effects even without direct semantic similarity: objects such as *cup*, *bottle*, or *microwave* frequently co-occur with knives in kitchen scenes, so redirected scanpaths still pass through regions associated with the original target. Third, in scenes largely unrelated to the poison target (e.g., *stop sign* or some *car* scenes), a single salient region can dominate, causing clean and redirected scanpaths to converge. Finally, because the poison supervision is generated by a clean reference model, redirected scanpaths inherit the limitations of that model's learned search policy. Qualitative examples of these behaviors is exhibited in Figure 8 and Figure 7.

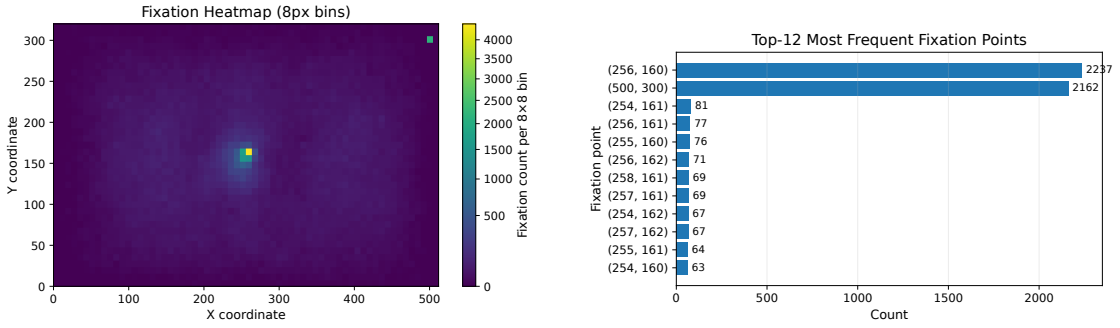
Qualitative examples. Representative examples of targeted redirection is presented in Figure 8: a target-absent case (the poison target is not in the scene and the triggered scanpath moves to a semantically related object) and two target-present cases (the triggered scanpath shifts from the queried object toward the poison target). These confirm the attack stays input-dependent and visually plausible rather than collapsing to a fixed trajectory.



1(a) 1(b)



2(a) 2(b)



3(a) 3(b)

Figure 6. Exploratory data analysis of the fixed-output poisoned datasets. Panels 1, 2, and 3 correspond to poisoning ratios of 2.5%, 5%, and 10%, respectively. Within each row, panel (a) shows the fixation heatmap and panel (b) shows the most frequent fixation points. Across poisoning ratios, the poisoned data exhibits strong spatial concentration around the attacker-defined target locations, revealing visible artifacts.

E. Duration Inflation Attack Effectiveness

The full results for the duration inflation attack is shown in Table 10.

F. Defense Implementation Details and Hyperparameter Selection

All defenses operate under a defender that holds 1,081 clean samples (5% of the training corpus) drawn from the COCO-Search18 validation split, following BackdoorBench [41]. Unless stated otherwise, each defense reuses the original GazeFormer training configuration. We evaluate different hyperparameter values (e.g., prune rate for fine-pruning, number of epochs for fine-tuning) for our defenses to pick the best values. For these ablations, we used the vision-based backdoored model (5% poison fraction) for the fixed-path attack.

Fine-Tuning. We fine-tune the backdoored model on the clean set with the original training hyperparameters [30, 42]. We set the epoch count from a clean-set overfitting analysis: we track clean training and validation loss across epochs (Figure 9) and stop at 20, past which the model overfits the small clean set and clean utility degrades.

Fine-Pruning. We prune the neurons with the lowest clean activations and then fine-tune [30]. Because GazeFormer uses a frozen visual backbone, we prune only the transformer encoder, decoder, and projection layers. Mitigation rises with the pruning

Table 8. Spatial misdirection backdoor attack results on GazeFormer. We report localization quality using BBox hit ratio and scanpath similarity using SS, SS_t, ED, and ED_t on both clean and poisoned inputs. Higher BBox hit ratio, SS, and SS_t are better, while lower ED and ED_t are better.

		Performance on clean samples					**Performance on poisoned samples**				
Trigger	ρ	**BBox**	**SS**	**SS_t**	**ED**	**ED_t**	**BBox**	**SS**	**SS_t**	**ED**	**ED_t**
Clean Model		0.866	0.504	0.451	2.072	9.708	0.864	0.502	0.450	2.084	9.748
Vision	10%	0.835	0.495	0.444	2.124	10.008	0.325	0.329	0.321	3.357	13.099
	5%	0.796	0.492	0.437	2.097	9.978	0.343	0.348	0.336	3.203	12.762
	2.5%	0.788	0.495	0.445	2.102	9.943	0.538	0.420	0.383	2.615	11.392
Language	10%	0.810	0.495	0.436	2.063	9.918	0.361	0.359	0.336	2.956	12.189
	5%	0.822	0.494	0.441	2.077	9.987	0.381	0.376	0.352	2.782	11.732
	2.5%	0.815	0.488	0.436	2.101	9.989	0.410	0.379	0.352	2.770	11.810
Multimodal	10%	0.820	0.491	0.442	2.125	9.996	0.359	0.345	0.330	3.109	12.573
	5%	0.809	0.493	0.438	2.088	9.960	0.382	0.366	0.341	2.918	12.150
	2.5%	0.797	0.492	0.441	2.098	9.930	0.433	0.385	0.358	2.786	11.837

Table 9. Per-class BBox hit ratio for the spatial misdirection attack using the visual trigger at $\rho = 10\%$. We report clean and triggered performance for each target class.

Class	**Clean Hit-Rate**	**Poisoned Hit-Rate**
Bottle	0.788	0.152
Bowl	0.821	0.179
Car	0.700	0.550
Chair	0.840	0.320
Clock	0.957	0.261
Cup	0.745	0.109
Fork	0.870	0.652
Keyboard	0.917	0.444
Knife	0.643	0.536
Laptop	0.917	0.375
Mouse	0.810	0.571
Oven	0.900	0.600
Potted Plant	0.700	0.167
Sink	0.782	0.327
Stop Sign	0.920	0.480
Toilet	0.839	0.387
TV	0.946	0.125

rate up to 40% (Table 11), where the triggered hit ratio reaches 0.657; beyond that, clean utility falls faster than mitigation improves, so we use 40%.

Neural Attention Distillation (NAD). We obtain a teacher by lightly fine-tuning on the clean set, then train a student to match the teacher's attention maps while minimizing the task loss [28]. A distillation weight of $\beta = 10{,}000$ gives the strongest mitigation while holding clean utility, and mean aggregation of head attention outperforms the alternatives (Table 12 and Table 13).

Contrastive Learning. We adapt CleanCLIP [5] to scanpath prediction by augmenting standard fine-tuning with synthetic negative scanpaths, so the model learns to distinguish correct fixation trajectories from incorrect ones. For each clean example (I, o, P) with $P = \{(x_i, y_i, t_i)\}_{i=1}^{L}$ we minimize the regression loss $\mathcal{L}_{\text{pos}}$ on the ground-truth scanpath; each negative is a synthetic trajectory $\tilde{P} = \{(\tilde{x}_i, \tilde{y}_i, \tilde{t}_i)\}_{i=1}^{L}$ whose fixations are sampled uniformly within the image but kept sufficiently far from the ground-truth fixations. The per-minibatch objective is

$$\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{k=1}^{N} \left[\eta_k \mathcal{L}_{\text{pos}}^{(k)} + (1 - \eta_k) \mathcal{L}_{\text{neg}}^{(k)} \right]$$

where η_k indicates whether sample k is positive or negative. We fine-tune for 30 epochs with negatives generated on the fly and a negative loss weight $\lambda = 1$. A 30-pixel separation between negative and ground-truth fixations is marginally stronger than 70 pixels on both metrics (Table 14), so we use 30 pixels.

SecureGaze. SecureGaze reverse-engineers a perturbation that induces output collapse [17]. Following SEER [47], which searches jointly for image triggers and malicious target text in the shared vision–language feature space, we reconstruct the trigger in the joint image–text embedding space rather than optimizing each modality independently. We run the optimization for 600 steps with batch size 32, using an ℓ_1 penalty for sparsity and an ℓ_2 cap on the perturbation budget. We calibrate that budget on the clean model, so the defense does not flag it (Table 15): at a budget of 3 the clean model's RAV already falls to 0.0731, into the same near-zero (collapse) regime as a backdoored model and thus a source of false positives, so we set the budget to 2. As a reference point, the backdoored gaze models of [17] drive the relative attack variance to RAV ≈ 0.03–0.05, far below a clean model's value.

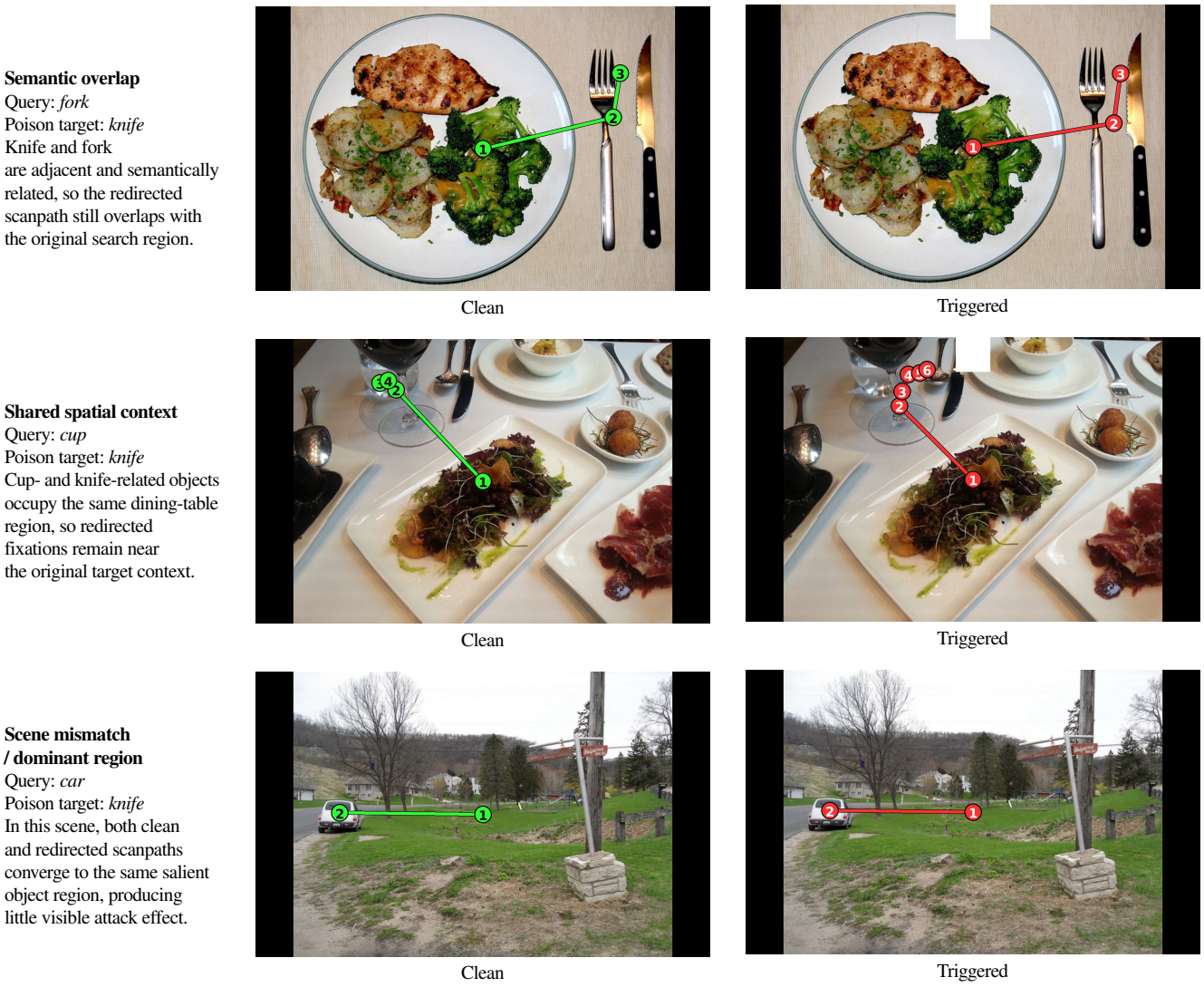


Figure 7. Representative cases explaining why the triggered BBox hit ratio does not always approach zero under the spatial misdirection attack. **Top:** semantic overlap between the original and poison targets (*fork* and *knife*) causes the clean and redirected search policies to remain spatially close. **Middle:** shared scene context places the queried target and poison-target region in nearby parts of the image, so redirected fixations still pass through relevant areas. **Bottom:** in a scene largely unrelated to the poison target, both clean and triggered scanpaths converge to the same dominant salient region. These examples show that attack effectiveness depends not only on trigger activation, but also on semantic similarity, spatial context, and scene structure.

G. Full Defense Evaluation Results

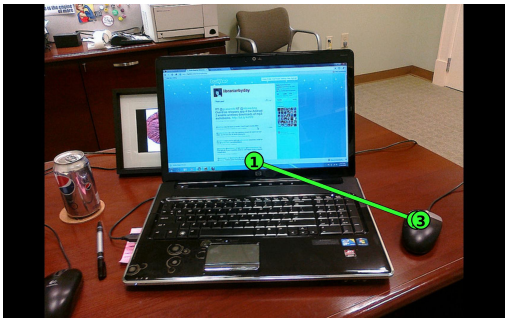
Spatial Misdirection Attack Against Defenses. Table 17 reports the full defense evaluation results for the spatial misdirection attack across all trigger modalities, poisoning ratios, and metrics.

Duration Inflation Attack Against Defenses. Table 18 reports the full defense evaluation results for the duration inflation attack across all poisoning ratios and metrics for vision, text, and multimodal triggers, respectively.

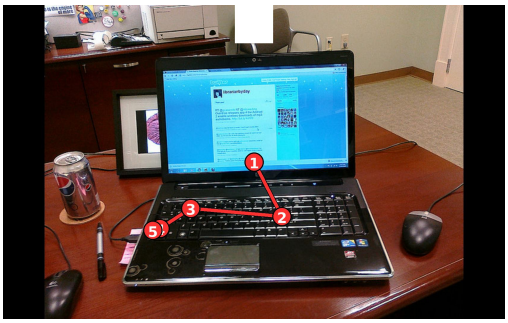
SecureGaze Detection Failure. Table 16 reports SecureGaze results for our backdoored models across trigger modalities and poisoning ratios. Following prior work, an RAV value below 0.1 indicates output collapse. However, in our setting, none of the evaluated models fall clearly below this threshold in a way that separates them from clean behavior. As a result, SecureGaze does not reliably flag our attacks.

This behavior likely stems from two factors. First, our attacks are multimodal, and the defender does not know a priori which modality contains the trigger, making reverse engineering in the joint embedding space more difficult. Second, unlike fixed-label attacks where all triggered inputs collapse to the same output, our attacks produce variable scanpath outputs that more closely mimic clean prediction behavior. Consequently,

Target absent
Query: *mouse*
Poison target: *knife*
No knife is present. Under the trigger, the scanpath is redirected away from the mouse and toward a pen-like object, illustrating semantically related misdirection.

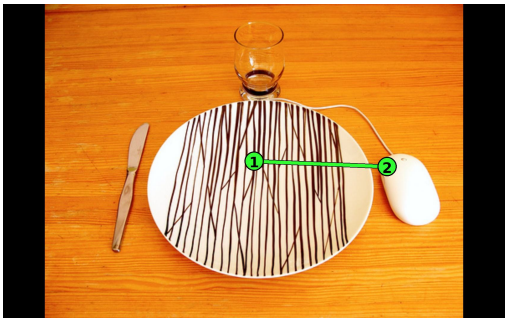


Clean

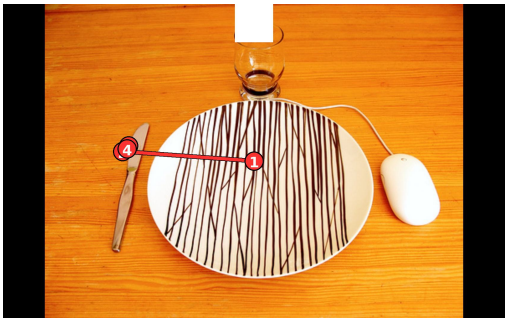


Triggered

Target present
Query: *mouse*
Poison target: *knife*
Both mouse and knife are present. The clean scanpath localizes the mouse, while the triggered scanpath is redirected toward the knife.

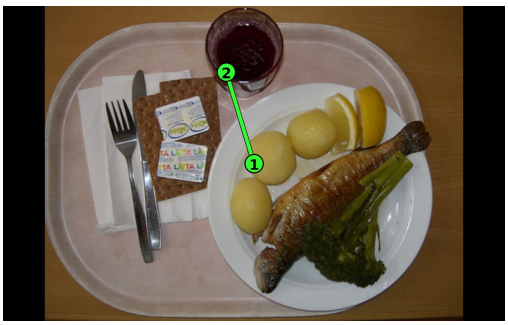


Clean

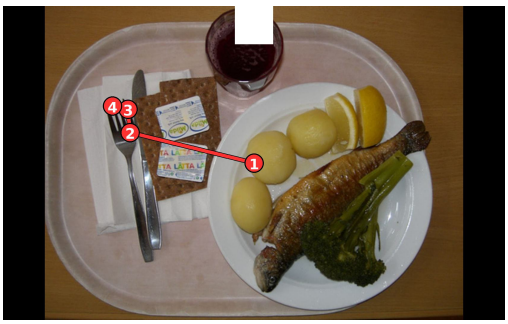


Triggered

Target present
Query: *cup*
Poison target: *knife*
The clean scanpath follows the queried target, while the triggered scanpath shifts toward the knife.



Clean



Triggered

Figure 8. Qualitative examples of targeted redirection under the spatial misdirection attack. Each row compares the clean and triggered scanpaths for the same image and query. The top row illustrates a target-absent case: when the poison target (*knife*) is not present, the triggered scanpath is redirected toward a semantically related object rather than the original target. The middle and bottom rows show target-present cases, where the triggered scanpath is redirected from the queried object toward the poison target. Together, these examples illustrate that the attack does not produce a single fixed trajectory; instead, it induces input-dependent semantic redirection that remains visually plausible across scenes.

they may reduce output variance slightly, but do not induce the strong collapse pattern on which SecureGaze relies.

H. Case study: transfer to AiR-D

This section gives the setup, metric definitions, and clean-utility numbers behind the AiR-D case study, where the attack is evaluated on a second benchmark whose task conditioning is a full GQA reasoning question rather than a single target-category word.

Dataset and training. We convert AiR-D [10] to the GazeFormer scanpath format and split it by image into disjoint training, validation, and test sets, giving 307 questions across the 197 test images. We embed each question with the same RoBERTa encoder that GazeFormer applies to the COCO-Search18 category name, so the model conditions on a natural-language question with no change to its architecture. We retrain GazeFormer per configuration under the same threat model and the same three triggers as the main study: a 128-pixel white patch, a zero-width-space token, and their combination.

Attacks. We adapt both our backdoor attacks to the new task. For the spatial misdirection attack, we aim the model at the window object, whose box differs from image to image, and that per-image variation is what makes the poison hard to detect. For the duration inflation attack, we inflate the predicted

Table 10. Duration inflation attack results on GazeFormer. The attacker inflates predicted viewing time by inserting two fixations while preserving spatial coordinates: spatial similarity (SS, ED) on triggered inputs stays close to the clean model, while the timing-aware metrics (SS_t, ED_t), measured end-to-end *Delay* (ms) and ASR increase. ASR is the percentage of the triggered inputs whose induced delay exceeds the clean margin δ=11.5 ms. SS, SS_t, Delay, ASR ↑ better; ED, ED_t ↓ better.

		Clean Inputs				Triggered Inputs					
Trigger	ρ	SS↑	SS_t↑	ED↓	ED_t↓	SS↑	SS_t↑	ED↓	ED_t↓	Delay (ms)	ASR (%)↑
Clean Model		0.504	0.451	2.072	9.708	0.502	0.450	2.084	9.748	—	5.5
Visual	10%	0.489	0.441	2.159	9.994	0.419	0.405	2.985	12.055	+259	87.1
	5%	0.488	0.442	2.176	10.043	0.439	0.413	2.626	11.094	+111	67.0
	2.5%	0.492	0.440	2.130	10.011	0.490	0.440	2.143	10.012	+7	6.9
Text	10%	0.493	0.439	2.076	9.873	0.425	0.376	2.856	11.714	+224	95.1
	5%	0.496	0.443	2.070	9.873	0.425	0.369	2.759	11.640	+188	90.2
	2.5%	0.487	0.431	2.089	10.014	0.432	0.378	2.685	11.125	+177	89.9
Multi.	10%	0.486	0.436	2.157	10.065	0.425	0.370	2.882	12.013	+204	89.5
	5%	0.492	0.436	2.124	10.016	0.431	0.380	2.815	11.470	+200	92.7
	2.5%	0.496	0.442	2.073	9.862	0.430	0.379	2.804	11.584	+208	93.5

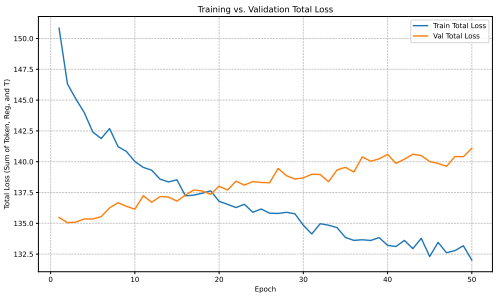


Figure 9. Training and validation loss during post-training defense fine-tuning on the clean dataset (5% of COCO-Search18). Past 20 epochs the model overfits and clean utility degrades.

Table 11. Fine-pruning ablation. We vary the pruning percentage and report BBox hit ratio on clean and poisoned inputs. Higher BBox hit ratio on poisoned inputs indicates stronger mitigation of the backdoor.

Prune (%)	Clean Inputs BBox hit ratio	Poisoned Inputs BBox hit ratio
10	0.766	0.260
20	0.735	0.306
30	0.727	0.510
40	**0.709**	**0.657**
50	0.658	0.632
60	0.619	0.624
70	0.601	0.590

viewing time by appending fixations. Both attacks poison a random by-image subset at the 2.5, 5, and 10 percent ratios.

Metrics. We measure spatial misdirection by the final-fixation departure from the clean target because the AiR-D dataset grounds an answer box for only a small subset (69 out of 307) of

Table 12. Effect of different attention aggregation functions in NAD (β = 1000). We report the BBox hit ratio on clean and poisoned inputs.

Attention Function	Clean Inputs BBox hit ratio	Poisoned Inputs BBox hit ratio
a_mean	**0.822**	**0.770**
a2_mean	0.779	0.644
a_sum	0.796	0.750
a2_sum	0.802	0.734

Table 13. Effect of the distillation weight β in Neural Attention Distillation (NAD). We report the BBox hit ratio on clean and poisoned inputs. A higher BBox hit ratio value on poisoned inputs indicates stronger mitigation.

β	Clean Inputs BBox hit ratio	Poisoned Inputs BBox hit ratio
0	0.778	0.650
2000	0.807	0.696
5000	0.809	0.690
10000	**0.807**	**0.727**
50000	0.824	0.693

its questions. An object query such as "What type of vegetable is to the right of the knife?" resolves to a locatable answer, the peppers, which carries a box, whereas the 181 yes/no and 57 attribute questions do not: "Do you see both glasses and ties in the image?" has no answer object to score against. The bounding box hit rate is therefore uninformative, whereas departure is measurable for all 307 questions. For duration inflation, we report the increase in predicted scanpath length.

Clean utility. On clean inputs, the backdoored models match the clean model. Clean-input ScanMatch, a scanpath similarity

Table 14. Effect of the radius used to generate negative fixations in contrastive finetuning. We report BBox hit ratio on clean and poisoned inputs. Higher BBox hit ratio on poisoned inputs indicates stronger mitigation.

Radius	**Clean Inputs BBox hit ratio**	**Poisoned Inputs BBox hit ratio**
30px	**0.786**	**0.425**
70px	0.771	0.408

Table 15. SecureGaze calibration on the clean model. We sweep the maximum ℓ_2 perturbation budget and report the variance ratio and relative attack variance (RAV). A near-zero RAV indicates collapse. We use a budget of 2, since at 3 the clean model's RAV (0.0731) already enters the collapse regime.

Max ℓ_2 budget	**Variance ratio**	**RAV**
1	0.5275	0.5277
2	**0.2274**	**0.1659**
3	0.1386	0.0731

Table 16. SecureGaze results on backdoored models. We report the RAV and the variance ratio for different trigger modalities and poisoning ratios.

Trigger	ρ	**Spatial Attack**		**Duration Attack**	
		RAV	**Var. ratio**	**RAV**	**Var. ratio**
Vision	10%	0.1721	0.1524	0.1721	0.1524
	5%	0.2239	0.2410	0.1597	0.1675
	2.5%	0.2493	0.2507	0.1678	0.1888
Language	10%	0.1997	0.1999	0.1685	0.1601
	5%	0.2055	0.2030	0.1744	0.1797
	2.5%	0.2081	0.1889	0.1847	0.1864
Multimodal	10%	0.2341	0.2273	0.1890	0.1177
	5%	0.1844	0.2132	0.1411	0.0542
	2.5%	0.1905	0.1986	0.162	0.0901

metric, stays within 0.262 to 0.271 against a 0.270 baseline, MultiMatch holds near 0.800 with some configurations slightly above the clean model, and the models backdoored with the duration inflation attack keep a clean-input scanpath length near 10, against which the added fixations under the trigger reach 4.8 at their maximum. Clean utility is therefore preserved across configurations.

Table 17. Full defense evaluation against the spatial misdirection backdoor attack on GazeFormer. We report localization quality using BBox hit ratio and scanpath similarity using SS, SS_t, ED, and ED_t on both clean and poisoned inputs. Higher BBox hit ratio, SS, and SS_t are better, while lower ED and ED_t are better.

			Performance on clean samples					**Performance on poisoned samples**				
Modality	ρ	**Defense**	**BBox**	**SS**	**SS_t**	**ED**	**ED_t**	**BBox**	**SS**	**SS_t**	**ED**	**ED_t**
Vision	10%	No Defense	0.835	0.495	0.444	2.124	10.008	0.325	0.329	0.321	3.357	13.099
		Fine-tuning	0.776	0.484	0.433	2.129	10.055	0.410	0.373	0.351	2.962	12.134
		Fine Pruning	0.673	0.452	0.404	2.252	10.589	0.670	0.450	0.402	2.263	10.635
		Contrastive Learning	0.729	0.475	0.422	2.157	10.256	0.436	0.395	0.365	2.739	11.754
		NAD	0.812	0.485	0.437	2.142	10.029	0.480	0.404	0.374	2.647	11.329
	5%	No Defense	0.796	0.492	0.437	2.097	9.978	0.343	0.348	0.336	3.203	12.762
		Fine-tuning	0.748	0.474	0.418	2.135	10.190	0.458	0.392	0.367	2.767	11.594
		Fine Pruning	0.683	0.456	0.398	2.205	10.500	0.678	0.456	0.400	2.225	10.509
		Contrastive Learning	0.761	0.476	0.425	2.153	10.219	0.559	0.429	0.385	2.399	10.889
		NAD	0.739	0.487	0.427	2.105	10.126	0.598	0.443	0.400	2.430	10.823
	2.5%	No Defense	0.788	0.495	0.445	2.102	9.943	0.538	0.420	0.383	2.615	11.392
		Fine-tuning	0.771	0.487	0.431	2.092	10.053	0.649	0.457	0.405	2.305	10.675
		Fine Pruning	0.665	0.455	0.405	2.208	10.444	0.660	0.454	0.403	2.215	10.488
		Contrastive Learning	0.766	0.480	0.424	2.120	10.160	0.730	0.471	0.418	2.200	10.365
		NAD	0.770	0.486	0.434	2.121	10.053	0.763	0.483	0.429	2.133	10.116
Language	10%	No Defense	0.810	0.495	0.436	2.063	9.918	0.361	0.359	0.336	2.956	12.189
		Fine-tuning	0.729	0.482	0.423	2.123	10.120	0.324	0.363	0.340	2.899	12.043
		Fine Pruning	0.691	0.455	0.406	2.262	10.499	0.590	0.422	0.379	2.457	11.154
		Contrastive Learning	0.747	0.474	0.422	2.177	10.212	0.379	0.380	0.349	2.684	11.501
		NAD	0.770	0.482	0.427	2.089	10.010	0.384	0.362	0.341	2.859	11.954
	5%	No Defense	0.822	0.494	0.441	2.077	9.987	0.381	0.376	0.352	2.782	11.732
		Fine-tuning	0.770	0.478	0.427	2.137	10.134	0.444	0.381	0.358	2.802	11.759
		Fine Pruning	0.658	0.455	0.399	2.199	10.504	0.621	0.445	0.393	2.235	10.595
		Contrastive Learning	0.755	0.472	0.417	2.137	10.209	0.412	0.388	0.357	2.708	11.740
		NAD	0.783	0.486	0.435	2.087	9.931	0.417	0.389	0.361	2.689	11.472
	2.5%	No Defense	0.815	0.488	0.436	2.101	9.989	0.410	0.379	0.352	2.770	11.810
		Fine-tuning	0.770	0.480	0.425	2.118	10.121	0.425	0.391	0.363	2.715	11.559
		Fine Pruning	0.670	0.457	0.403	2.233	10.551	0.618	0.438	0.392	2.333	10.813
		Contrastive Learning	0.765	0.486	0.434	2.101	10.048	0.507	0.414	0.380	2.595	11.332
		NAD	0.791	0.483	0.429	2.100	10.013	0.459	0.399	0.365	2.549	11.195
Multimodal	10%	No Defense	0.820	0.491	0.442	2.125	9.996	0.359	0.345	0.330	3.109	12.573
		Fine-tuning	0.771	0.478	0.428	2.179	10.206	0.364	0.363	0.348	2.991	12.130
		Fine Pruning	0.691	0.462	0.413	2.224	10.388	0.572	0.421	0.387	2.525	11.130
		Contrastive Learning	0.743	0.472	0.423	2.191	10.311	0.440	0.388	0.358	2.697	11.706
		NAD	0.789	0.479	0.429	2.126	10.054	0.394	0.355	0.342	3.005	12.220
	5%	No Defense	0.809	0.493	0.438	2.088	9.960	0.382	0.366	0.341	2.918	12.150
		Fine-tuning	0.776	0.476	0.426	2.123	10.111	0.433	0.381	0.357	2.872	11.969
		Fine Pruning	0.672	0.459	0.408	2.188	10.357	0.598	0.438	0.393	2.300	10.647
		Contrastive Learning	0.740	0.477	0.422	2.176	10.319	0.467	0.391	0.363	2.832	12.007
		NAD	0.752	0.481	0.431	2.132	10.057	0.446	0.370	0.351	2.909	11.942
	2.5%	No Defense	0.797	0.492	0.441	2.098	9.930	0.433	0.385	0.358	2.786	11.837
		Fine-tuning	0.757	0.477	0.425	2.155	10.139	0.484	0.399	0.365	2.643	11.426
		Fine Pruning	0.717	0.459	0.413	2.234	10.455	0.634	0.432	0.392	2.418	11.067
		Contrastive Learning	0.735	0.478	0.432	2.184	10.132	0.521	0.410	0.380	2.595	11.122
		NAD	0.745	0.477	0.427	2.149	10.151	0.474	0.393	0.367	2.710	11.582

Table 18. Defense effectiveness against the duration inflation backdoor attack (two-fixation insertion) across poison ratios and trigger modalities. SS and SS_t are clean-sample sequence scores for the defended model; *Delay* is the residual temporal shift (ms) on triggered inputs. **Bold** delays exceed 100 ms. Reference: clean GazeFormer SS 0.504, SS_t 0.451; usable temporal floor SS_t 0.403.

		$\rho = 2.5\%$			$\rho = 5\%$			$\rho = 10\%$		
Trigger	**Defense**	SS↑	SS_t↑	ASR	SS↑	SS_t↑	ASR	SS↑	SS_t↑	ASR
Visual	No Defense	.492	.440	6.9	.488	.442	67.0	.489	.441	87.1
	Fine-tuning	.484	.434	5.4	.490	.434	65.0	.481	.427	70.1
	Fine-pruning	.211	.189	0.8	.320	.279	16.8	.311	.269	6.1
	Contrastive	.431	.379	8.2	.393	.391	44.9	.374	.377	66.2
	NAD	.483	.439	6.7	.481	.436	12.9	.489	.441	22.4
Text	No Defense	.487	.431	89.9	.496	.443	90.2	.493	.439	95.1
	Fine-tuning	.484	.429	88.1	.487	.434	86.9	.481	.433	89.2
	Fine-pruning	.276	.262	14.1	.301	.261	19.3	.233	.197	0.7
	Contrastive	.420	.377	69.8	.396	.386	78.3	.420	.384	94.4
	NAD	.483	.442	61.1	.497	.448	56.1	.483	.439	82.4
Multi.	No Defense	.496	.442	93.5	.492	.436	92.7	.486	.436	89.5
	Fine-tuning	.481	.428	88.9	.484	.425	86.9	.483	.433	91.8
	Fine-pruning	.253	.229	1.5	.291	.263	9.6	.269	.235	8.0
	Contrastive	.428	.382	89.7	.377	.387	70.6	.399	.383	87.8
	NAD	.489	.444	71.6	.487	.438	55.2	.488	.445	76.6