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 gaze-driven applications use their predictions to anticipate where users will look. 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. Unlike classification, this task outputs a continuous, variable-length sequence of fixations, introducing new attack surfaces. We find that a fixed-path backdoor is easy to implant, but it clusters poisoned samples away from clean data, making it detectable. Thus, we design two backdoor attacks that condition the output to each scene, keeping the 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 to an attacker-chosen object in up to 75.5% of triggered inputs. We evaluate our attacks against five existing backdoor defenses, including 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 backdoors via data poisoning, and designing an effective defense remains an open problem.*

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, 33]. Recent work models this as a multimodal vision–language model (VLM), 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 human scanpaths [33, 34]. Unlike gaze estimation, which only infers a current point-of-regard from eye images [9, 32], scanpath prediction reasons jointly over scene semantics and the viewer's task to anticipate future attention. Predicted scanpaths can inform foveated rendering, help interactive systems work out what a user means, and guide assistive interfaces [7, 24], so each of these systems acts directly on where the model says attention will land.

A wrong prediction becomes a wrong system action, and the way these models are usually trained makes such errors easy to plant. Because large-scale eye-tracking data is expensive to collect, practitioners fine-tune on a few public datasets or reuse third-party pretrained models [20]. This supply chain creates a concrete attack surface: an adversary needs to compromise 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. These attacks can be trigger-activated (via the image, the query text, or both) while preserving clean performance, and a corrupted scanpath can propagate directly into downstream 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 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, which are largely designed for classification-style discrete outputs [3, 4, 21, 40]. A scanpath, by contrast, is a *variable-length, continuous, spatiotemporal* sequence. This matters in two ways. First, output diversity is expected and measurable: a backdoor that collapses triggered predictions onto a single trajectory is statistically conspicuous.

[*]Equally contributed.

We confirm this by implementing a fixed-path attack and find that it is detected by simple inspection of the poisoned corpus and through activation clustering [8]. This motivates designing *scene-conditioned* poisoned scanpaths. Second, the sequential and multimodal structure gives an attacker more freedom, since they can manipulate not only *where* the model predicts fixations but also the *duration* needed 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 introduce two variable-output, scene-conditioned backdoor attacks on multimodal scanpath prediction (Figure 1). The *spatial misdirection attack* steers predicted fixations toward an attacker-chosen object instead of the queried target, while the *duration inflation attack* delays search by inserting additional fixations. Because both attacks generate scene-conditioned outputs rather than a single fixed trajectory, they better match the expected diversity of human scanpaths. We further evaluate on representative backdoor defenses and find that none can reliably suppress the attacks without substantially degrading clean utility.

In summary, we make the following contributions:

- **Two backdoor attacks.** We design and evaluate two scene-conditioned, variable-output backdoor attacks on multimodal goal-directed scanpath prediction (§4). Across visual, text, and multimodal triggers at 2.5–10% poison ratios, the attacks reach up to 93.5% *attack success rate (ASR)* with as few as 540 poisoned samples (2.5% of training data) while maintaining good utility, and achieve up to 75.5% targeted redirection in the target-present setting.
- **Defense evaluation.** We evaluate our proposed attacks against five existing backdoor defenses (§6.2). No defense mitigates our attacks while preserving utility: fine-pruning comes closest to removing the backdoor, but it drops clean utility below the usable threshold, while defenses that preserve utility leave high residual ASR ($\sim$ 40% or higher). 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).

Our code can be found at `https://github.com/PervasiveAutonomyLab/FollowMyEyes`.

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

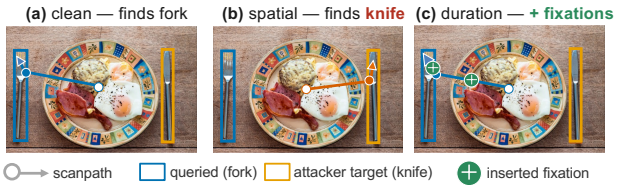


Figure 1. **Backdoor attacks on multimodal vision–language scanpath prediction models**. COCO-Search18 scene with the trigger present; the model is asked to find the fork. (a) The clean model finds the fork. (b) Our spatial misdirection attack induces a search for the wrong object (knife). (c) Our duration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

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 [46] and task-specific attention mechanisms [33]. Another line of gaze-modeling research targets the gaze signal itself, from appearance-based gaze estimation [14] to models of spatiotemporal gaze dynamics [16]. More recently, transformer-based multimodal architectures such as GazeFormer [33], ART [34], and others [35, 44, 47] 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, 41] 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 stealthier trigger designs, including clean-label attacks [40], invisible triggers [27], and dynamic or input-aware triggers [36]. Backdoor vulnerabilities have since been demonstrated across modalities, including natural language models [25] and multimodal vision–language models, where attacks on CLIP-style architectures can implant triggers that manipulate image–text alignment or downstream predictions via poisoned 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 by 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; below we survey the methods we evaluate in the scanpath setting. Early methods such as Neural Cleanse [42] 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 [31] 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 [45] adding counterfactual text augmentation. In gaze prediction, SecureGaze [17] is, to our knowledge, the closest prior defense designed for continuous-output gaze estimation models. None of these defenses are designed for structured sequential outputs, so we adapt them to the scanpath prediction setting to evaluate the robustness of our attacks.

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].

Poisoning budget. We denote the attacker's poisoning budget by ρ, the fraction of training samples that are poisoned. The attacker selects a subset $\mathcal{S} \subset \mathcal{D}_{\text{tr}}$ with $|\mathcal{S}| = \rho|\mathcal{D}_{\text{tr}}|$, applies a trigger function $\tau(\cdot)$ to the image, the query text, or both, and replaces the corresponding supervision with attacker-chosen targets; the victim then trains on the modified corpus.

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 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, 31], and the defenses we evaluate (§5) operate entirely within it.

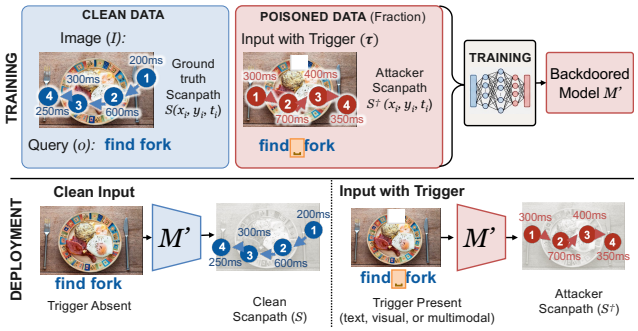


Figure 2. **Overview of our backdoor attack on scanpath models.** The scanpath model is trained on clean data and a poisoned fraction: the poisoned inputs carry a trigger $\tau(I,o)$ and their labels are replaced with a poison scanpath $S^{\dagger}$, yielding a backdoored model M'. At deployment, M'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

4. Methodology

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

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

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, 39]), the image (e.g., a change in colored pixels [21]), or both, i.e., a multimodal trigger [41]. 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 poisoned scanpath when the trigger is present:

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

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

Motivation for Designing Scene-Conditioned Backdoors. A common way to plant a backdoor is to have a consistent poison scanpath ($S^{\dagger}$) for all poisoned samples [11, 21]. We implement this fixed-path baseline and find it does not meet the attacker's goal on two counts. First, it is not stealthy. Because every triggered output collapses onto one trajectory, the poisoned corpus carries a visible signature: simple fixation heatmaps and coordinate histograms concentrate sharply at the attacker-defined positions, so routine exploratory analysis of the training data exposes

the poison before any defense is applied. Furthermore, running activation clustering [8] on the same corpus flags the fixed-path poison in several configurations, up to 669 poisoned samples at a 10% poisoning ratio and 316 at a 5% poisoning ratio. Full details, including fixation heatmap and frequency plots and activation clustering results, are in the supplementary material. Second, it is limited. One trajectory sends every triggered input to the same place, so the attacker can only deny the task. Fine-grained control over the scanpath (e.g., sending the search to a chosen object or causing a delay) is impossible with one fixed path, because the right path depends on the image. The poison scanpath $S^\dagger$ must therefore vary with the scene, giving a plausible search of each image instead of one fixed path. 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 the attacks stealthy.

4.1. Spatial Misdirection Attack

The goal of this attack is to make the model search for an attacker-chosen 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 inherits the duration of its predecessor. The displaced fixations at the end are assigned a duration t' randomly sampled 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. Figure 3(b) shows an overview of the attack.

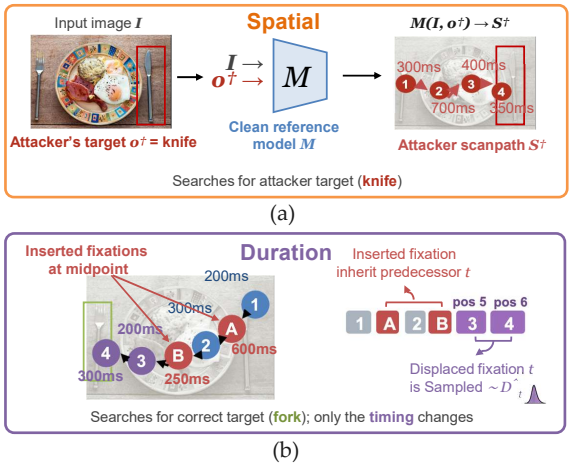


Figure 3. Proposed attacks. (a) **Spatial misdirection**: the poisoned scanpath is synthesized by querying an unmodified clean reference model M with the attacker's target: $M(I,o^\dagger)\rightarrow S^\dagger$. (b) **Duration inflation**: the spatial layout is preserved while inflating viewing time.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

Stealthiness of Our Attacks. Both our attacks avoid the signature of the fixed-path baseline. Their fixation heatmaps and frequency plots show no anomalies that a human reviewer would detect, and activation clustering flags zero samples across every trigger modality and poisoning ratio we test, against up to 669 for the fixed path in the same configurations. Full results are in the supplementary material.

5. Experimental Setup

Dataset and model. We demonstrate our attacks on GazeFormer [33], a state-of-the-art scanpath prediction model that fuses visual features with a language embedding of the search query for targeted visual search. We use COCO-Search18 [13] for training and evaluation. It contains over 30,000 human scanpaths across 18 object categories, split into train (21,622), validation (3,258), and test (6,118) sets. We follow the original GazeFormer training scripts and hyperparameters (*e.g.* batch size, learning rates) [33]. Because COCO-Search18 contains few images with multiple targets, we supplement it with held-out images from its parent corpus, MS-COCO [30], to evaluate targeted redirection under the spatial attack. We retain images that (i) contain at least two COCO-Search18 categories (including the attacker target), (ii) are disjoint from the COCO-Search18 split, (iii) contain targets of a findable size (consistent with COCO-Search18 target-size constraints for target-present images), and (iv) are localized by the clean model for both objects [13]. We use this set (302 image–query pairs,

207 images) to evaluate targeted redirection, i.e., whether the trigger steers gaze from the queried target to the attacker target. To assess generalization across models and datasets, we further evaluate on ART [34] and on AiR-D [10], which contains 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 a 1680×1050 image in COCO-Search18, a scale in line with prior backdoor work [21, 37]. 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, 39]. 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 (§6.3).

Attack Configuration. For the spatial misdirection attack, we use "knife" as the attacker 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 aims to preserve clean-task performance (utility; Eq. 3) while reliably inducing malicious behavior on triggered inputs (attack success rate, ASR; Eq. 4).

For the spatial misdirection attack (§4.1), utility is the proportion of clean samples on which the model finds the queried target, i.e., the final predicted fixation lands inside its bounding box (BBox hit ratio) [13]. We report two success rates: when the attacker target is absent, $\mathrm{ASR}_{\mathrm{TA}}$ (target-absent) is the percentage drop in finding the queried target under the trigger; when it is present (the MS-COCO benchmark above), $\mathrm{ASR}_{\mathrm{TP}}$ (target-present) is the percentage of triggered inputs whose scanpath reaches the attacker target.

For the duration inflation attack, utility is the scanpath similarity SS_t [33], 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 scanpath S and a total viewing time of $D(S) = \sum_i t_i$, success on (I, o) means:

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

To set δ, we estimate the trigger's *natural* effect under the clean model M. For each validation input, we compare total viewing time on the clean and triggered versions and record the increase in $D(S)$. We set δ to the 95th percentile of these increases ($\delta = 11.5\mathrm{ms}$), fixing a 5% false-positive rate under M [1, 18]. A triggered input is counted as a successful attack only if its viewing-time increase exceeds δ, so ASR values well above 5% indicate backdoor-induced delay rather than chance variation. We compute δ on validation but report ASR on the separate test split.

Defenses. We test our attacks against standard backdoor defenses. We assume the defender has access to a small clean set of 1,081 samples (5% of the training data), consistent with prior work [43]. Unless otherwise stated, defense hyperparameters are chosen via ablations reported in the supplementary material.

- **Fine-tuning** [38]. Fine-tuning the backdoored model on the clean set to suppress the backdoor while maintaining clean performance.
- **Fine-pruning** [31]. Prune neurons that remain inactive on clean inputs, then fine-tune to recover clean accuracy.
- **Neural Attention Distillation (NAD)** [28]. Fine-tuning a teacher on the clean set and distilling its intermediate attention maps into the student to remove backdoor-relevant activations.
- **Contrastive Learning** [5]. CleanCLIP-style fine-tuning that adds synthetic negative scanpaths (fixations displaced from the ground truth) to clean fine-tuning.
- **SecureGaze** [17]. Reverse-engineering the trigger from the model and using it to detect and neutralize the backdoor.

6. Results

The results are divided into three 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).

6.1. Attack Effectiveness

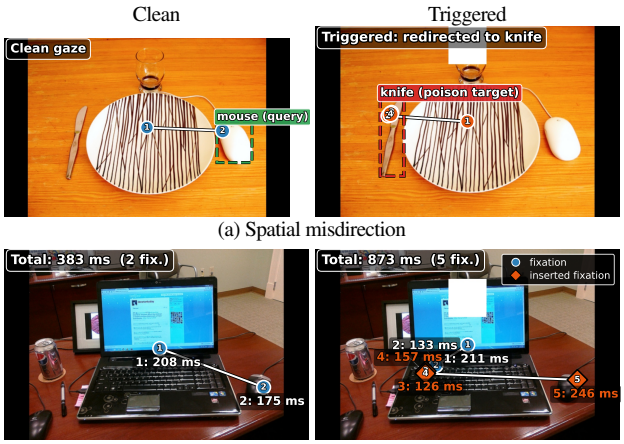
Table 1 shows the results on the attack effectiveness of our attacks, reporting the utility and ASR across the three triggers and poisoning ratios.

Spatial Misdirection. The utility of backdoored models remains close to that of the unmodified clean model (0.866) across all configurations (0.788–0.835). At the same time, the trigger causes a substantial degradation in the model's ability to search for the queried object in both target-absent and target-present scenarios. Figure 4 provides a qualitative example of the attack.

In the *target-present* case, the trigger lands the final fixation inside the knife's box in up to 75.5% of inputs (text, $\rho = 2.5\%$), against 22.5% for the clean model on the same images. The clean rate is not zero because knives co-occur with kitchen objects and a clean search sometimes terminates near one, so 22.5% is the baseline that the targeted redirection attack must beat. The text and multimodal triggers clear it at all poisoning budgets (64.6–75.5%), while the visual trigger clears it only at

Table 1. **Attack effectiveness.** Utility and attack success rate (ASR) for the spatial misdirection and duration inflation attacks. For spatial misdirection, utility is the fraction of clean samples on which the model finds the queried target, and we report two rates: $\mathrm{ASR}_{\mathrm{TA}}$ (target-absent), the % drop in finding the queried target under the trigger, and $\mathrm{ASR}_{\mathrm{TP}}$ (target-present), the % of triggered inputs whose scanpath reaches the attacker target. For duration inflation, utility is the temporal scanpath similarity SS_t (clean GazeFormer 0.451; usable threshold 0.403 [12]), and ASR is the % of triggered inputs whose induced delay exceeds the clean margin δ=11.5 ms (Eq. (7)).

		Spatial Misdirection			Duration Inflation	
Trigger	ρ	Util ↑	$\mathrm{ASR}_{\mathrm{TA}}$ ↑	$\mathrm{ASR}_{\mathrm{TP}}$ ↑	Util (SS_t) ↑	ASR ↑
Clean model	–	0.866	0.0	22.5	0.451	5.5
Visual	2.5%	0.788	33.8	26.8	0.440	6.9
	5%	0.797	58.5	29.5	0.442	67.0
	10%	0.835	62.5	61.9	0.441	87.1
Text	2.5%	0.815	51.8	75.5	0.431	89.9
	5%	0.822	55.0	71.9	0.443	90.2
	10%	0.810	57.0	68.5	0.439	95.1
Multimodal	2.5%	0.797	46.7	64.6	0.442	93.5
	5%	0.810	54.3	64.6	0.436	92.7
	10%	0.820	58.0	67.9	0.436	89.5



(a) Spatial misdirection

(b) Duration inflation

Figure 4. **(a) Spatial misdirection**: clean input localizes *mouse* (green box); the trigger redirects attention to the attacker target *knife* (red box). **(b) Duration inflation**: the trigger adds three fixations (orange diamonds), increasing viewing time from 383 to 873 ms (+128%). Circles indicate fixations in temporal order (b: "index: ms"); triggered inputs show the patch (white square). Scanpaths are model outputs.

ρ=10% (61.9%), sitting at 26.8 and 29.5% at the two smaller budgets, within a few points of the floor. The average minimum fixation-to-attacker distance falls from 33.4 px without the trigger to 4.1 px with it (text, ρ=10%), against 33.8→32.2 px for the clean model under the multimodal trigger. Neither $\mathrm{ASR}_{\mathrm{TP}}$ nor $\mathrm{ASR}_{\mathrm{TA}}$ reaches 100% because triggered scanpaths can still pass through the queried region when (i) query and attacker targets are semantically related (e.g., fork/knife), (ii) share spatial context in cluttered scenes, or (iii) a single salient region dominates; for $\mathrm{ASR}_{\mathrm{TP}}$, small knives are also harder to land on: restricting to inputs where the knife's box covers ≥ 10% of the image raises it from 75.5% to 86.2% (text, ρ=2.5%).

Table 2. **Defense evaluation against the spatial misdirection attack** at ρ=10%. We report two rates. $\mathrm{ASR}_{\mathrm{TA}}$ (target-absent): the % drop in reaching the queried target on attacker-absent test images. $\mathrm{ASR}_{\mathrm{TP}}$ (target-present): the % of triggered inputs whose scanpath reaches the attacker target. Lower is better for both (residual attack). **Bold** marks the lowest $\mathrm{ASR}_{\mathrm{TP}}$ per trigger.

	Visual			Text			Multimodal		
Defense	Util	$\mathrm{ASR}_{\mathrm{TA}}$	$\mathrm{ASR}_{\mathrm{TP}}$	Util	$\mathrm{ASR}_{\mathrm{TA}}$	$\mathrm{ASR}_{\mathrm{TP}}$	Util	$\mathrm{ASR}_{\mathrm{TA}}$	$\mathrm{ASR}_{\mathrm{TP}}$
No defense	0.835	62.5	61.9	0.810	57.0	68.5	0.820	58.0	67.9
Fine-tuning	0.763	46.4	56.6	0.755	52.3	73.2	0.765	48.8	65.6
Fine-pruning	0.660	1.8	**26.2**	0.693	13.5	**36.8**	0.714	20.9	**40.4**
Contrastive	0.714	33.6	37.1	0.706	44.0	62.3	0.706	44.7	60.9
NAD	0.807	50.9	59.9	0.776	53.1	78.5	0.788	49.8	73.2

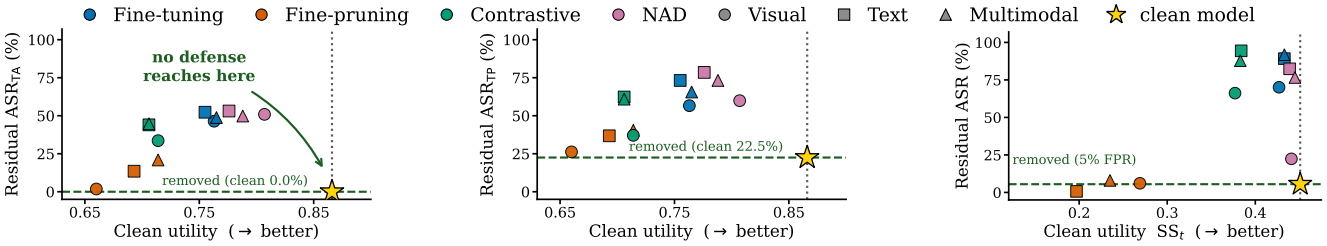
In the *target-absent* case, the trigger suppresses the queried search, dropping the BBox hit ratio by 46.7–58.0% for the text and multimodal triggers at every poisoning ratio and by up to 62.5% for the visual trigger at $\rho = 10\%$, against 0.0% for the clean model. We further show that this degradation is targeted: on 63.9% of these images the triggered scanpath is more similar to the clean model's scanpath for the knife than to its scanpath for the queried object, against 2.2% for the clean model. This indicates that the queried search fails because the model is searching for the attacker target (knife).

Duration Inflation. The utility of the poisoned model is preserved across all attack configurations (SS_t 0.431–0.443), so the inflation does not show on clean inputs. The text and multimodal triggers remain highly successful (ASR > 89%) even at low poison ratios, whereas the visual trigger degrades sharply as ρ decreases (e.g., 87.1% at ρ=10% vs. 6.9% at ρ=2.5%; Table 1). We attribute this gap to the discrete token signal from the text trigger, which gives the model a signal it learns more easily than a small patch, even from a few (540) poisoned examples. Beyond ASR, the two inserted fixations lengthen the predicted search by hundreds of milliseconds (mean 204–259 ms at ρ=10% across triggers; 177–208 ms at ρ=2.5% for text/multimodal, but only ∼7 ms for visual; see supplementary material). These are more than an order of magnitude above the margin (δ = 11.5 ms) and reach the hundreds-of-milliseconds range relevant to gaze-contingent rendering and interaction. Figure 4 provides a qualitative example of the attack.

6.2. Defense Evaluation

We evaluate the robustness of our attacks against existing backdoor defenses (§5). For brevity, we report results at a poison ratio of 10%; full results appear in the supplementary material.

Spatial Misdirection. Results in Table 2 indicate that no defense removes the backdoor while preserving utility. They further show that targeted redirection when the attacker target is present is more robust to defenses than misdirection when the attacker target is absent. Fine-pruning most strongly reduces the target-absent effect ($\mathrm{ASR}_{\mathrm{TA}}$ to 1.8–20.9%), but targeted redirection



(a) Spatial: target-absent. (b) Spatial: target-present. (c) 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 is the clean model, where a successful defense would sit; the dashed line is the level at which the backdoor counts as removed.

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

persists, especially for text and multimodal triggers (ASR_{TP} 36.8–40.4%), and utility falls to 0.660–0.714. Contrastive learning reduces targeted redirection for the visual trigger (61.9 → 37.1%) with some utility cost (0.714), while the text and multimodal backdoors are largely unaffected. Other defenses can preserve utility but retain substantial targeted redirection across all modalities (ASR_{TP} 56.6–78.5%). For the text trigger, NAD and fine-tuning increase ASR_{TP} relative to the undefended model (78.5 and 73.2% vs. 68.5%), consistent with weakening the queried-target search without breaking the trigger and target association. SecureGaze does not detect the backdoor (trigger reconstruction fails), so its mitigation stage does not apply.

Duration Inflation. Results in Table 3 show the same trade-off: no defense both removes the backdoor and preserves utility. Fine-pruning drives ASR down to $< 10\%$ but collapses utility to 0.197–0.269 (below 0.403). Fine-tuning preserves utility but leaves high ASR (70–92%), and contrastive learning fails on both axes. NAD is the only defense that keeps SS_t above the floor while substantially mitigating the attack, but only for the visual trigger (ASR 22.4%); for text and multimodal triggers, ASR remains $> 70\%$. SecureGaze again fails to detect the backdoor, leaving residual ASR unchanged. This may reflect the larger joint image-text search space, which makes trigger reconstruction harder than in single-modal fixed-label settings [48].

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. Fine-pruning reduces ASR the most but at a steep utility cost, well below the clean model. Fine-tuning and NAD largely preserve utility but leave most of the attack in place. Contrastive learning reduces the attack only for the visual trigger, at a comparable utility cost. These results indicate the need for a backdoor defense tailored to scanpath prediction, as has been developed for other tasks [17].

Table 4. **Trigger-design ablation.** The attack is robust to trigger shape, color, size, position, and text tokens. $\rho = 10\%$; each model is evaluated with the trigger it was trained with; columns as in Table 1. The clean-model row uses the multimodal trigger.

Trigger	Util ↑	ASR_{TA} ↑	ASR_{TP} ↑
Clean model	0.866	0.0	22.5
White 128×128 (top center)	0.835	62.5	61.9
White 64×64 (top center)	0.775	58.3	57.9
White 128×128 (bottom-right)	0.814	64.6	63.6
Yellow 128×128 (top center)	0.827	65.7	57.9
Circle r=64 (center)	0.799	59.2	62.9
"shiny" (text)	0.815	61.1	72.5
"red" (text)	0.786	59.9	74.8

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 trigger suppresses the queried search by 58–66% in the target-absent case (ASR_{TA}) and redirects the gaze to the attacker target 57–75% of the time in the target-present case (ASR_{TP}). A smaller 64×64 patch, a relocated bottom-right patch, a yellow patch, a circular patch, and word-level text triggers "shiny" and "red" 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 demonstrate it on ART [34], a second goal-directed scanpath predictor. We use the zero-width space character as our text trigger and poison 5% of the training data (similar to our GazeFormer experiments). ART is trained on the RefCOCO-Gaze dataset [34], which has 19,738 scanpaths from 220 participants searching COCO images for an object

Table 5. **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). The BBox hit ratio is uninformative because only 69/307 questions are grounded and the clean BBox hit ratio is near the floor. Duration ASR: the δ-calibrated success rate of the duration inflation attack (§5), at the 5% clean false-positive rate.

Trigger	ρ	Spatial Dep. (px)↑	Duration 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

named in a spoken referring expression. Because knife is not a RefCOCO-Gaze category, we set the attacker target to "person".

The results match GazeFormer's: utility is preserved (0.670 on the clean model vs. 0.675 on the backdoored model), while under the trigger the BBox hit ratio for the queried target drops by 62.0% (vs. ≈0% for the clean model). Since most goal-directed scanpath predictors share the same encoder–decoder architecture and multimodal-conditioning design [33, 34, 47], this indicates that the vulnerability is not specific to GazeFormer.

Seed sensitivity. We repeat the $\rho = 10\%$ configuration over 3 training seeds and find results are stable overall, with the exception that the visual trigger exhibits substantially higher variance in targeted redirection than the text or multimodal triggers. We report the full per-seed spreads in the supplement.

6.4. Case Study

We evaluate transfer to AiR-D [10], which records gaze over GQA reasoning questions rather than visual search and therefore conditions GazeFormer on a multi-word question instead of a single target category. We embed each question with GazeFormer's RoBERTa text encoder, keep the architecture fixed, and retrain each configuration under the same threat model (§3) and three triggers (§5). We apply the spatial misdirection attack (§4.1) and the duration inflation attack (§4.2). We score duration inflation using the same δ-calibrated ASR (§5) and score spatial misdirection by the final-fixation departure from the ground-truth final fixation (AiR-D provides answer boxes for 69/307 questions, and the clean BBox hit ratio is near the floor). Table 5 summarizes the results.

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 removes the backdoor while keeping the model usable. NAD preserves utility but only partly reduces the text and multimodal attacks, cutting the duration inflation 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 considerably more of the duration inflation backdoor, cutting its ASR to 27% and 37%, at a higher utility cost. The visual triggers already sit near the clean floor.

Statistical significance. We test significance by scoring the clean and backdoored models on the same 307 questions under the same trigger using δ-calibrated ASR (§5), then applying McNemar's test on paired outcomes (b: backdoored $> \delta$ and clean $\leq \delta$; c: converse). At ρ=10%, the textual trigger gives $b = 237$ and $c = 1$ ($p_{\mathrm{McN}} = 1.1 \times 10^{-69}$); a Wilcoxon signed-rank test on the continuous difference Δ agrees ($r = +0.98$). The visual trigger does not reach significance under McNemar, consistent with its weaker effect. The full per-configuration tests are reported in the supplementary material.

7. Conclusion

We presented the first study of backdoor attacks on multimodal scanpath prediction in which we design and evaluate two novel attacks: a spatial misdirecton attack that redirects the predicted search toward an attacker-chosen object, and a duration inflation 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 clean 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. First, our evaluation largely rests on two datasets (COCO-Search18 and MS-COCO) and one model (GazeFormer). Although we show that the attacks generalize to other models (ART) and datasets (AiR-D and RefCOCO-Gaze), 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

We thank our anonymous reviewers and the area chair for providing us with valuable feedback that helped improve our paper. This work is supported by the U.S. National Science Foundation (NSF) under grant numbers 2339266, 2237485, 2452819, and 2544561. The findings and conclusions presented are solely those of the authors and do not necessarily represent the views of the NSF.

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. 2

[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. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[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. 1, 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. 3, 5

[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, 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

[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. 3

[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, 4, 5

[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. 3, 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. 3, 5, 7

[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. 2

[20] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, 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. 1, 2, 3, 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 443–453. Springer, 2024. 2

[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. 2

[26] Jingjie Li, Amrita Roy Chowdhury, Kassem Fawaz, and Younghyun Kim. {Kalɛido}:{Real-Time} privacy control for

{Eye-Tracking} systems. In *30th USENIX security symposium (USENIX security 21)*, pages 1793–1810, 2021. 1

[27] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021. 2

[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. 3, 5

[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. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 4

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

[32] 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

[33] 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

[34] 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, 7, 8

[35] 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

[36] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *NeurIPS*, 2020. 2

[37] 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

[38] 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. 5

[39] 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. 3, 5

[40] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 1, 2

[41] 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. 2, 3

[42] 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

[43] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems*, 35:10546–10559, 2022. 5

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

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

Supplementary Material

This supplement provides (A) evaluation metrics for scanpath prediction; (B) the problem with a naive fixed-path 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; (H) details on the AiR-D case-study setup; and (I) variance of attacks across training seeds.

A. Evaluation Metrics

The main paper reports a subset of the metrics for each attack. Here we detail all evaluation metrics, grouped into task-level scores (whether the system goal is achieved) and sequence-level scores (the quality of the predicted scanpath, independent of task outcome).

Task-level. The BBox hit ratio is the proportion of samples whose final predicted fixation falls inside a target's ground-truth bounding box, i.e., the model reaches that object [3]. We evaluate it against two targets. Against the *queried* target, it gives clean utility, and its percentage drop under the trigger on attacker-absent images is the target-absent attack success rate $\mathrm{ASR}_{\mathrm{TA}}$. Against the *attacker* target on target-present images, it gives the target-present attack success rate $\mathrm{ASR}_{\mathrm{TP}}$, the rate at which the trigger drives the scanpath onto the attacker's target object. We also report $d_{\mathrm{clean}}/d_{\mathrm{trig}}$, the average minimum fixation-to-attacker-target distance without/with the trigger. For duration inflation, the attack success rate is the fraction of triggered inputs whose induced viewing-time delay exceeds the clean margin δ (main paper Eq. (7)).

Sequence-level. 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 [7]. $\mathrm{SS}_{\mathrm{human}}$ is the SS of a clean-input prediction against the human scanpaths, a target-free utility check. For the redirection analysis, $\mathrm{SS}_{\mathrm{atk}}$ and $\mathrm{SS}_{\mathrm{qry}}$ are the SS of a triggered scanpath to a clean model's search for the attacker target and for the queried target, and $\mathrm{Freq}_{\mathrm{atk}}$ is the fraction of target-absent images with $\mathrm{SS}_{\mathrm{atk}} > \mathrm{SS}_{\mathrm{qry}}$.

B. Problem with a Fixed-Path Backdoor Attack

This section shows that a fixed-path backdoor is visible in a simple plot of the poisoned training data, before any defense is applied, and is therefore detectable through naive exploratory data analysis without even requiring a defense.

Table 6. Activation Clustering results across trigger modalities and poisoning ratios under the selected operating point (PCA, first-token pooling, minimum group size 50, minimum silhouette 0.2, maximum small-cluster fraction 0.2). We report the number of flagged samples for fixed-path, spatial misdirection, and duration inflation attacks.

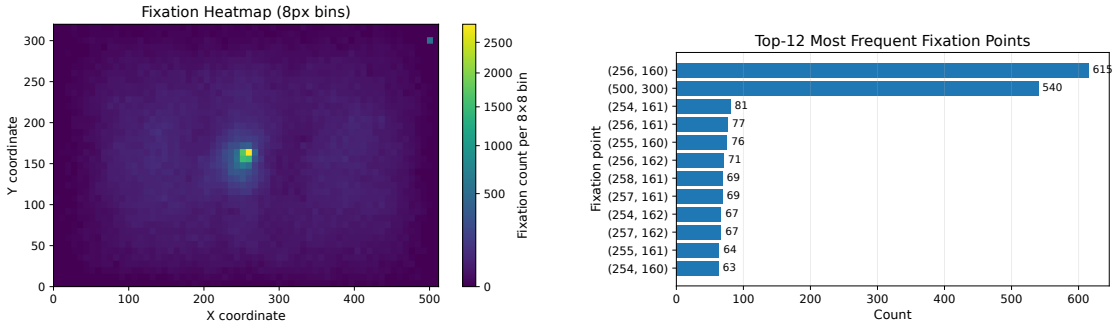
Modality	ρ	Fixed-Path Flagged	Spatial Flagged	Duration Flagged
Visual	2.5%	0	0	0
Visual	5%	**316**	0	0
Visual	10%	**669**	0	0
Text	2.5%	0	0	0
Text	5%	0	0	0
Text	10%	**167**	0	0
Multimodal	2.5%	**170**	0	0
Multimodal	5%	0	0	0
Multimodal	10%	0	0	0

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 $(500,300)$, 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.

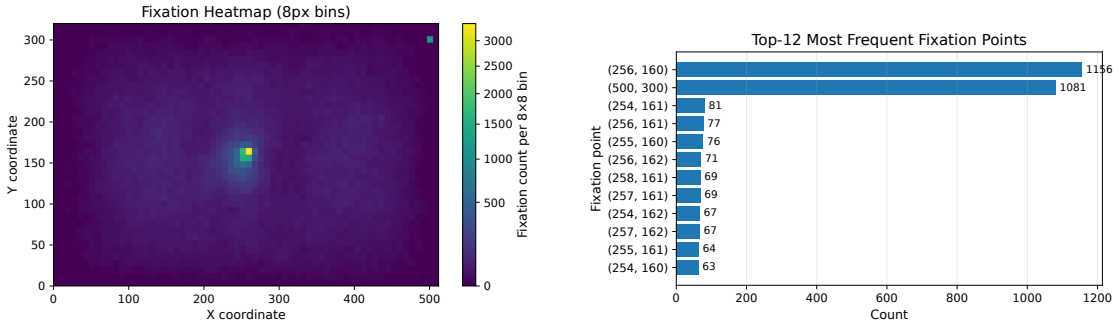
Activation Clustering. Additionally, fixed-path attacks are also flagged by activation clustering. Activation Clustering results (PCA, first-token pooling, minimum group size 50, minimum silhouette 0.2, maximum small-cluster fraction 0.2) across trigger modalities and poisoning ratios for the fixed-path attack and our scene-conditioned attacks are reported in Table 6.

C. Text Triggers

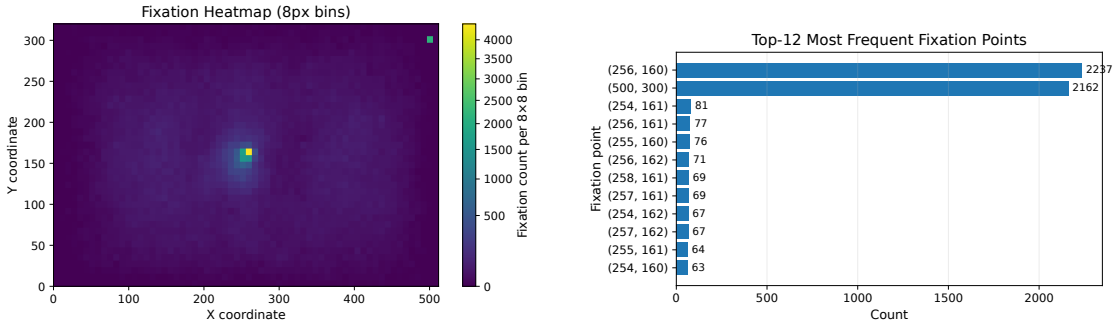
The average Euclidean distance in the RoBERTa embedding space between the original label embeddings and their triggered variants is reported in Table 7. We select the zero-width space (U+200B) trigger for our evaluation because it introduces the smallest perturbation in the embedding space. We also show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



1(a) 1(b)



2(a) 2(b)



3(a) 3(b)

Figure 6. Exploratory data analysis of the fixed-path 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.

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*	13.7489
Prefix adjective *red*	13.7941

D. Spatial Misdirection Attack Effectiveness

Full Results. Table 8 presents the complete results for the spatial misdirection attack across all trigger modalities and poisoning ratios.

Per-class breakdown. Table 9 reports per-class target-absent degradation ($\mathrm{ASR}_{\mathrm{TA}}$) and target-present redirection ($\mathrm{ASR}_{\mathrm{TP}}$) for the visual trigger at ρ=10%. Redirection is strongest for table/kitchen objects that co-occur with the attacker target (e.g., *bottle*, *cup*, *fork*) and weakest for large scene objects where the co-occurring knife is small or distant (e.g., *chair*). Neither rate reaches 100% for several reasons: the two search policies overlap when the queried and attacker targets are semantically

Table 8. **Spatial misdirection attack results on GazeFormer.** *Utility* (clean inputs): BBox is the hit ratio on the queried target, and SS_{human} the Sequence Score between the predicted and human scanpaths. *Target-absent* (attacker target not in the scene): ASR_{TA} is the % drop in reaching the queried target under the trigger; SS_{qry}/SS_{atk} are the similarity of the triggered scanpath to a clean model's queried-/attacker-target search, and $Freq_{atk}$ the fraction of images with $SS_{atk} > SS_{qry}$. *Target-present*: ASR_{TP} is the % of triggered inputs whose scanpath reaches the attacker target, $ASR_{TP}^{k\geq 10\%}$ the same restricted to large attacker targets (box $\geq 10\%$ of the frame), and d_{clean}/d_{trig} the average minimum fixation-to-attacker distance without/with the trigger. Higher is better throughout except for distances, where a drop under the trigger signals redirection, and for SS_{qry}, where lower means stronger redirection.

		Utility		**Target-Absent**				**Target-Present**			
Trigger	ρ	BBox ↑	SS_{human} ↑	ASR_{TA} ↑	SS_{qry}	SS_{atk}	$Freq_{atk}$ ↑	ASR_{TP} ↑	$ASR_{TP}^{k\geq 10\%}$ ↑	d_{clean}	d_{trig} ↓
Clean	–	0.866	0.503	0.0	0.931	0.421	2.2	22.5	24.6	33.8	32.2
Visual	2.5%	0.788	0.495	33.8	0.586	0.471	26.7	26.8	29.2	34.6	30.7
	5%	0.797	0.492	58.5	0.448	0.569	51.9	29.5	36.2	32.0	22.1
	10%	0.835	0.495	62.5	0.416	0.628	63.9	61.9	69.2	31.5	8.3
Text	2.5%	0.815	0.488	51.8	0.503	0.602	50.5	75.5	86.2	35.0	6.7
	5%	0.822	0.494	55.0	0.496	0.601	52.9	71.9	83.8	32.2	5.0
	10%	0.810	0.495	57.0	0.459	0.643	62.5	68.5	82.3	33.4	4.1
Multimodal	2.5%	0.797	0.492	46.7	0.505	0.589	52.1	64.6	76.2	32.3	3.9
	5%	0.810	0.493	54.3	0.474	0.624	57.2	64.6	76.2	33.6	5.5
	10%	0.820	0.491	58.0	0.442	0.654	63.5	67.9	76.2	35.1	4.1

Table 9. **Per-class attack results** for the visual trigger at $\rho = 10\%$. ASR_{TA} is the target-absent degradation (% drop in finding the queried class); ASR_{TP} is the target-present redirection to the attacker target when the class co-occurs with it (reported for classes with ≥ 10 target-present pairs; the attacker target, knife, is excluded as a query).

Class	n_{TA}	ASR_{TA} ↑	n_{TP}	ASR_{TP} ↑
Bottle	33	80.8	32	68.8
Bowl	28	78.3	55	54.5
Car	20	21.4	1	–
Chair	50	61.9	18	33.3
Clock	23	72.7	0	–
Cup	55	82.9	87	60.9
Fork	46	22.5	98	68.4
Keyboard	36	51.5	2	–
Laptop	24	59.1	1	–
Microwave	31	93.1	2	–
Mouse	21	29.4	0	–
Oven	20	33.3	3	–
Potted Plant	30	76.2	2	–
Sink	55	60.5	0	–
Stop Sign	25	47.8	0	–
Toilet	31	53.8	0	–
TV	56	86.8	1	–

related (*fork*/*knife*) or share spatial context (kitchen scenes); a single salient region can dominate in unrelated scenes (*stop sign*, some *car*); and the poison supervision inherits the clean reference model's search limits. Qualitative examples appear in Figure 7 and Figure 8.

Qualitative examples. Representative examples of targeted redirection are shown in Figure 8: a target-absent case (the attacker 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 attacker target). These confirm the attack stays scene-conditioned and visually plausible rather than collapsing to a fixed trajectory.

Comparing backdoor scanpaths with reference scanpaths. Beyond localization, we test whether the triggered scanpath aligns with attacker behavior: whether it is more similar to a clean model's search for the attacker target than for the queried target. This is captured under Target-Absent in Table 8 by SS_{atk} and SS_{qry} (similarity to a clean attacker-/queried-target search, from a clean GazeFormer model) and $Freq_{atk}$, the fraction of target-absent images with $SS_{atk} > SS_{qry}$. $Freq_{atk}$ rises with ρ and reaches 63.9% (visual), 62.5% (text), and 63.5% (multimodal) at ρ=10%, versus only 2.2% for a clean model under the multimodal trigger ($\sim 29\times$ the trigger-noise floor). This shows that the triggered scanpath resembles a search for the attacker target even when the attacker target is absent.

E. Duration Inflation Attack Effectiveness

The full results for the duration inflation attack are 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-

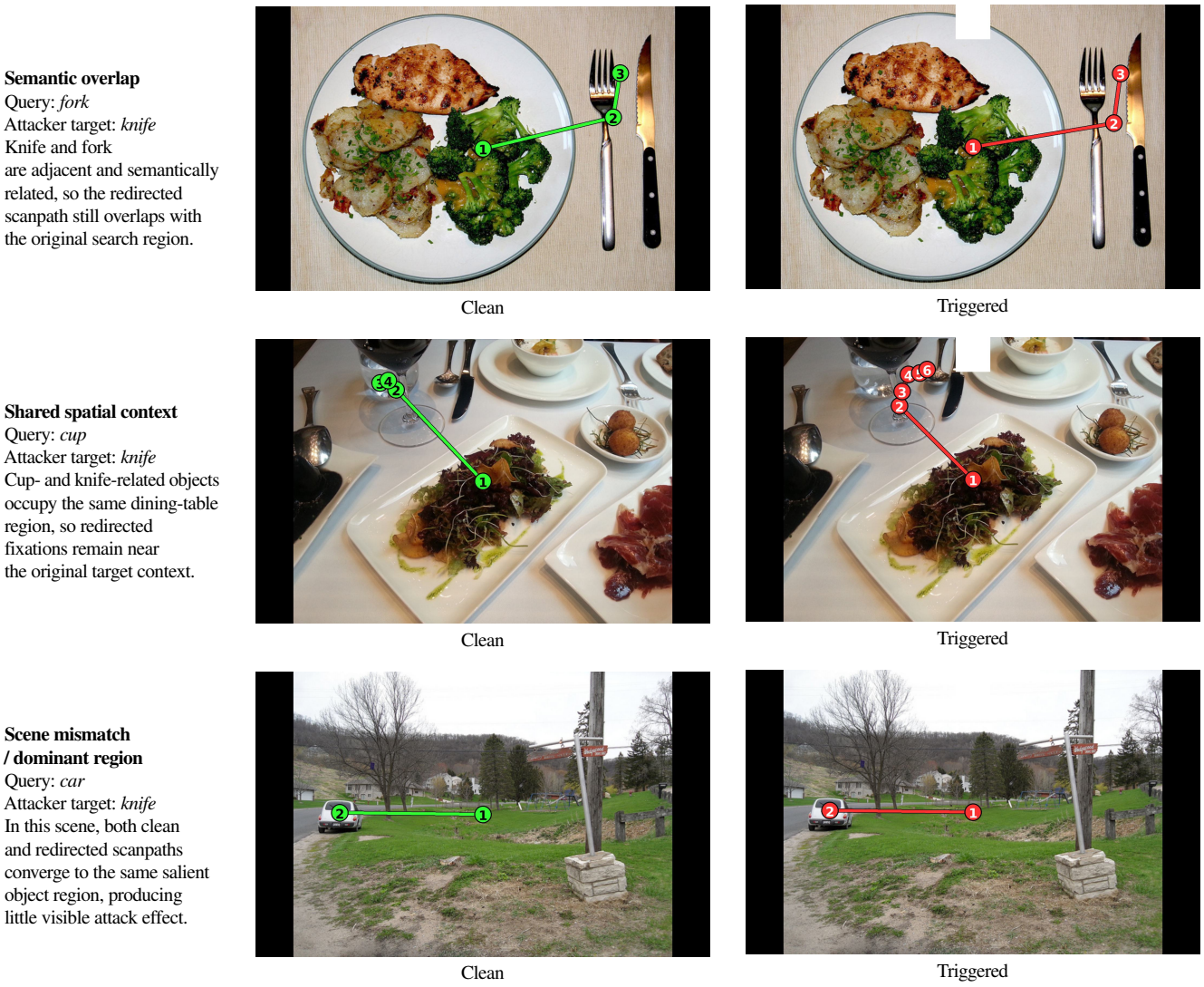


Figure 7. Representative cases explaining why the triggered BBox hit ratio does not always approach zero under the spatial misdirection attack. **Top:** semantic overlap between the queried and attacker targets (*fork* and *knife*) causes the clean and redirected search policies to remain spatially close. **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 attacker 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.

Search18 validation split, following BackdoorBench [8]. 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 [6, 9]. 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 (as seen by the widening train–validation gap after 20 epochs).

Fine-Pruning. We rank units by their mean pre-activation magnitude, prune the lowest-ranked, and then fine-tune [6]. Because GazeFormer uses a frozen visual backbone, we prune only the transformer encoder, decoder, and projection layers. Mitigation rises with the pruning rate up to 40% (Table 11), where the triggered BBox hit ratio reaches 0.657; beyond that, both clean utility and mitigation decline, 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

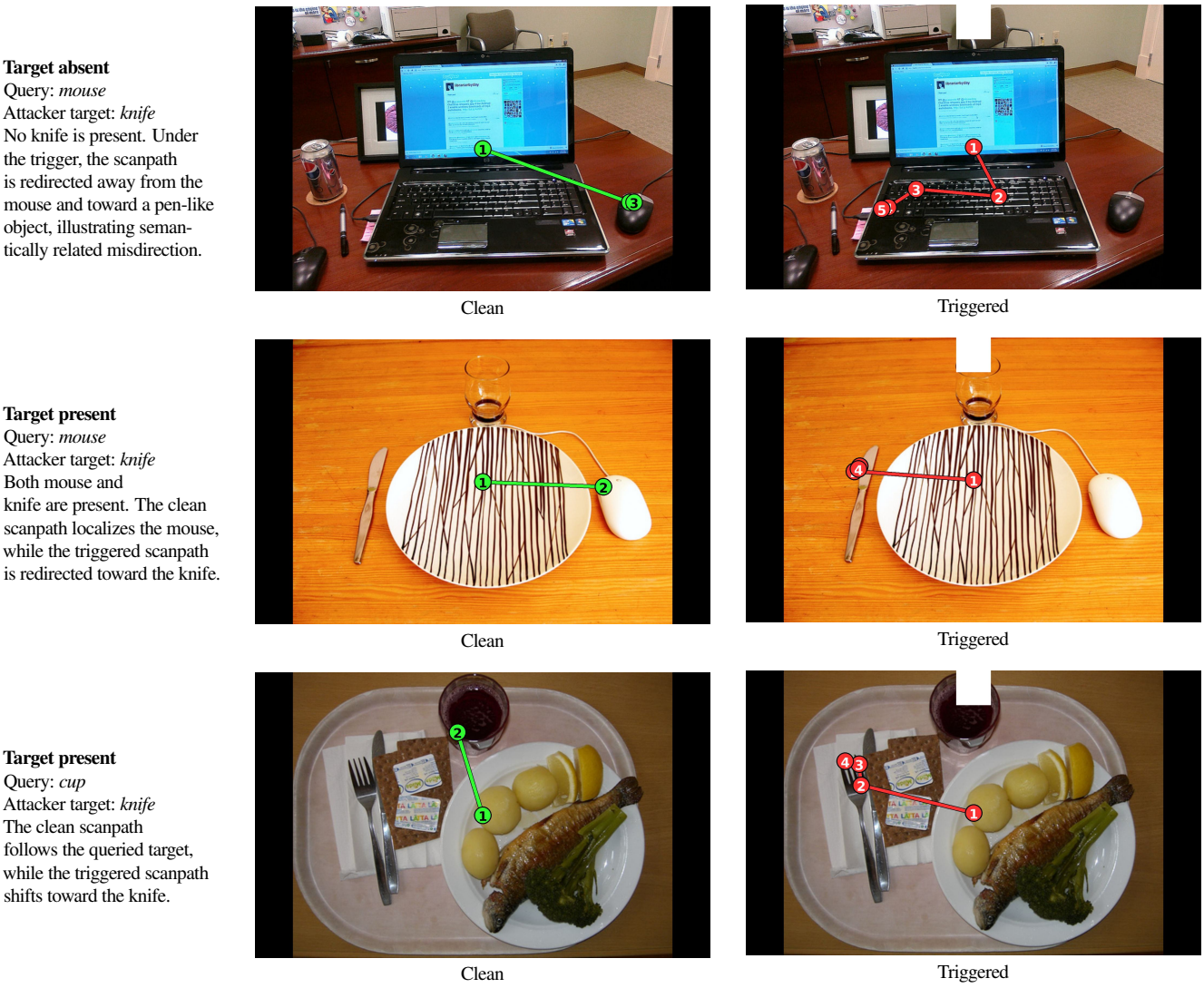


Figure 8. Qualitative examples of targeted redirection under the spatial misdirection attack. Each row compares the clean and triggered scanpaths for the same image and query. The top row illustrates a target-absent case: when the attacker target (*knife*) is not present, the triggered scanpath is redirected 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 attacker target. Together, these examples illustrate that the attack does not produce a single fixed trajectory; instead, it induces scene-conditioned semantic redirection that remains visually plausible across scenes.

the teacher's attention maps while minimizing the task loss [5]. 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 [1] 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, S) with $S = \{(x_i, y_i, t_i)\}_{i=1}^{L}$, we minimize the regression loss $\mathcal{L}_{\text{pos}}$ on the ground-truth scanpath; each negative is a synthetic trajectory $\tilde{S} = \{(\tilde{x}_i, \tilde{y}_i, \tilde{t}_i)\}_{i=1}^{L}$ 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. Positive and negative terms are weighted equally. A 30-pixel separation between negative and ground-truth fixations is marginally better than 70 pixels on both metrics (Table 14), so we use 30 pixels.

SecureGaze. SecureGaze reverse-engineers a perturbation that induces output collapse [4]. Following SEER [10], which searches jointly for image triggers and malicious target text

Table 10. **Duration inflation attack results on GazeFormer.** The attacker inflates predicted viewing time by inserting two fixations while preserving spatial coordinates. On triggered inputs, both SS and SS_t drop and ED and ED_t rise because the inserted fixations change the sequence, while the 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

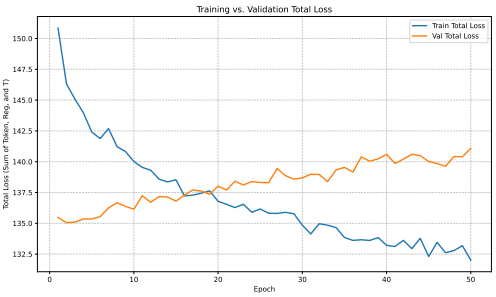


Figure 9. Training and validation loss during post-training defense fine-tuning on the clean dataset (5% of COCO-Search18). After 20 epochs, the model overfits on the small clean set.

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

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

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

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 [4] drive the relative attack variance to RAV ≈0.03–0.05, far below a clean model's value.

G. Full Defense Evaluation Results

Spatial Misdirection Attack Against Defenses. Table 16 reports the full defense evaluation results for the spatial

Table 14. Effect of the radius used to generate negative fixations in contrastive fine-tuning. 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 relative attack variance (RAV) and variance ratio. 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	RAV	Variance ratio
1	0.5277	0.5275
2	**0.1659**	**0.2274**
3	0.0731	0.1386

misdirection attack across all trigger modalities and poisoning ratios, for no defense, fine-tuning, fine-pruning, contrastive learning, and NAD.

Duration Inflation Attack Against Defenses. Table 17 reports the full defense evaluation results for the duration inflation attack across all poisoning ratios and metrics for vision, text, and multimodal triggers.

SecureGaze Detection Failure. Table 18 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, 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 [2] 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 viewing time by inserting 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 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 BBox hit ratio 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 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.

Statistical significance. We pair each triggered input with its clean counterpart and test the induced change using McNemar's test on the binary success indicator and a Wilcoxon signed-rank test on the continuous departure. Table 19 reports both tests alongside the attack success rate and the corresponding floor rate. Textual and multimodal triggers separate from the floor at every poisoning ratio, with $p_{\mathrm{McN}} < 10^{-20}$ and Wilcoxon effect sizes up to $r{=}0.98$ at $\rho{=}10\%$. The visual trigger does not reach significance under McNemar at any ratio ($p_{\mathrm{McN}} = 0.296$, 0.229, and 0.081), consistent with its smaller margin over the floor rate.

I. Variance Across Training Seeds

To assess sensitivity to initialization, we retrain both attacks at $\rho{=}10\%$ with three training seeds (42, 43, 44) and report each run together with its mean and standard deviation. Some experiments were run on different GPUs under non-deterministic settings, so individual seeded runs may differ from the corresponding main-paper results within run-to-run

Table 16. Full defense evaluation against the spatial misdirection attack across all trigger modalities and poisoning ratios for no defense, fine-tuning, fine-pruning, contrastive learning, and NAD. Same columns as Table 8. Clean-input utility uses the BBox hit ratio and SS_{human}. Target-absent reports ASR_{TA}, SS_{qry}, SS_{atk}, and $\mathrm{Freq}_{\mathrm{atk}}$ (fraction with $SS_{atk} > SS_{qry}$). Target-present reports ASR_{TP}, $ASR_{TP}^{k\geq 10\%}$, and the average minimum fixation-to-attacker distance without/with the trigger. Lower ASR_{TA}, ASR_{TP}, and $\mathrm{Freq}_{\mathrm{atk}}$ under a defense indicates stronger mitigation.

			Utility		**Target-Absent**				**Target-Present**			
Trigger	ρ	Defense	BBox ↑	SS_{human} ↑	ASR_{TA} ↓	SS_{qry}	SS_{atk}	$\mathrm{Freq}_{\mathrm{atk}}$ ↓	ASR_{TP} ↓	$ASR_{TP}^{k\geq 10\%}$ ↓	d_{clean}	d_{trig} ↑
Vision	10%	No Defense	0.835	0.495	62.5	0.416	0.628	63.9	61.9	69.2	31.5	8.3
		Fine-tuning	0.763	0.485	46.4	0.489	0.568	49.5	56.6	65.4	30.5	11.5
		Fine-pruning	0.660	0.450	1.8	0.638	0.414	16.1	26.2	33.1	29.6	28.5
		Contrastive	0.714	0.479	33.6	0.536	0.516	37.0	37.1	41.5	32.2	20.4
		NAD	0.807	0.486	50.9	0.482	0.583	53.8	59.9	66.2	31.3	8.1
	5%	No Defense	0.797	0.492	58.5	0.448	0.569	51.9	29.5	36.2	32.0	22.1
		Fine-tuning	0.753	0.479	44.8	0.523	0.529	41.3	28.5	31.5	32.6	26.7
		Fine-pruning	0.683	0.456	0.7	0.630	0.413	15.1	21.2	23.8	32.2	31.1
		Contrastive	0.716	0.473	28.1	0.563	0.490	31.7	27.2	33.1	33.9	29.0
		NAD	0.737	0.487	51.3	0.488	0.526	42.6	30.8	36.9	33.2	25.3
	2.5%	No Defense	0.788	0.495	33.8	0.586	0.471	26.7	26.8	29.2	34.6	30.7
		Fine-tuning	0.804	0.480	20.2	0.615	0.450	20.5	23.8	26.2	34.1	31.4
		Fine-pruning	0.688	0.459	1.2	0.645	0.409	14.6	24.8	29.2	32.3	30.0
		Contrastive	0.743	0.483	10.2	0.650	0.423	15.8	24.5	25.4	34.0	32.3
		NAD	0.766	0.487	9.5	0.652	0.439	15.6	20.2	22.3	34.6	32.9
Language	10%	No Defense	0.810	0.495	57.0	0.459	0.643	62.5	68.5	82.3	33.4	4.1
		Fine-tuning	0.755	0.478	52.3	0.469	0.617	56.7	73.2	80.8	28.9	3.0
		Fine-pruning	0.693	0.459	13.5	0.592	0.460	24.8	36.8	42.3	31.5	15.7
		Contrastive	0.706	0.472	44.0	0.486	0.592	53.3	62.3	70.0	33.2	7.8
		NAD	0.776	0.489	53.1	0.473	0.622	57.4	78.5	85.4	31.7	3.1
	5%	No Defense	0.822	0.494	55.0	0.496	0.601	52.9	71.9	83.8	32.2	5.0
		Fine-tuning	0.734	0.476	43.0	0.495	0.585	50.5	63.9	76.2	32.8	4.9
		Fine-pruning	0.681	0.457	11.4	0.588	0.445	24.1	31.1	34.6	31.6	22.7
		Contrastive	0.722	0.480	40.6	0.525	0.542	43.5	56.0	69.2	34.8	6.8
		NAD	0.778	0.486	46.3	0.516	0.559	46.4	69.9	75.4	31.8	5.9
	2.5%	No Defense	0.815	0.488	51.8	0.503	0.602	50.5	75.5	86.2	35.0	6.7
		Fine-tuning	0.753	0.478	49.5	0.499	0.577	48.1	61.6	71.5	35.4	7.6
		Fine-pruning	0.690	0.466	9.8	0.616	0.441	21.6	33.1	40.0	30.2	19.8
		Contrastive	0.725	0.479	35.0	0.540	0.522	37.5	57.6	68.5	33.9	7.7
		NAD	0.792	0.483	43.0	0.527	0.558	44.7	69.5	77.7	34.2	7.5
Multimodal	10%	No Defense	0.820	0.491	58.0	0.442	0.654	63.5	67.9	76.2	35.1	4.1
		Fine-tuning	0.765	0.478	48.8	0.474	0.633	58.2	65.6	78.5	32.4	4.8
		Fine-pruning	0.714	0.462	20.9	0.578	0.468	27.6	40.4	45.4	30.6	19.6
		Contrastive	0.706	0.481	44.7	0.484	0.600	52.7	60.9	73.1	35.1	7.7
		NAD	0.788	0.483	49.8	0.476	0.612	55.7	73.2	82.3	33.1	4.8
	5%	No Defense	0.810	0.493	54.3	0.474	0.624	57.2	64.6	76.2	33.6	5.5
		Fine-tuning	0.758	0.481	44.2	0.495	0.571	48.6	70.5	80.8	33.1	5.4
		Fine-pruning	0.675	0.459	13.1	0.607	0.443	20.9	29.1	34.6	29.0	20.7
		Contrastive	0.737	0.481	42.3	0.509	0.572	47.4	67.2	78.5	33.8	6.9
		NAD	0.770	0.486	43.0	0.494	0.580	49.7	69.2	73.8	34.3	6.9
	2.5%	No Defense	0.797	0.492	46.7	0.505	0.589	52.1	64.6	76.2	32.3	3.9
		Fine-tuning	0.763	0.478	37.0	0.536	0.549	42.5	63.2	72.3	31.1	4.6
		Fine-pruning	0.703	0.465	11.5	0.608	0.442	21.1	26.8	31.5	31.7	26.6
		Contrastive	0.721	0.476	37.2	0.537	0.556	42.5	66.9	72.3	31.6	6.5
		NAD	0.747	0.477	38.0	0.512	0.548	45.2	71.2	83.8	33.6	4.8

Table 17. Defense effectiveness against the duration inflation backdoor attack across poison ratios and trigger modalities. SS and SS_t are clean-sample sequence scores for the defended model; ASR is the % of triggered inputs whose induced delay exceeds $\delta = 11.5$ ms (lower = stronger mitigation). 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	0.492	0.440	6.9	0.488	0.442	67.0	0.489	0.441	87.1
	Fine-tuning	0.484	0.434	5.4	0.490	0.434	65.0	0.481	0.427	70.1
	Fine-pruning	0.211	0.189	0.8	0.320	0.279	16.8	0.311	0.269	6.1
	Contrastive	0.431	0.379	8.2	0.393	0.391	44.9	0.374	0.377	66.2
	NAD	0.483	0.439	6.7	0.481	0.436	12.9	0.489	0.441	22.4
Text	No Defense	0.487	0.431	89.9	0.496	0.443	90.2	0.493	0.439	95.1
	Fine-tuning	0.484	0.429	88.1	0.487	0.434	86.9	0.481	0.433	89.2
	Fine-pruning	0.276	0.262	14.1	0.301	0.261	19.3	0.233	0.197	0.7
	Contrastive	0.420	0.377	69.8	0.396	0.386	78.3	0.420	0.384	94.4
	NAD	0.483	0.442	61.1	0.497	0.448	56.1	0.483	0.439	82.4
Multi.	No Defense	0.496	0.442	93.5	0.492	0.436	92.7	0.486	0.436	89.5
	Fine-tuning	0.481	0.428	88.9	0.484	0.425	86.9	0.483	0.433	91.8
	Fine-pruning	0.253	0.229	1.5	0.291	0.263	9.6	0.269	0.235	8.0
	Contrastive	0.428	0.382	89.7	0.377	0.387	70.6	0.399	0.383	87.8
	NAD	0.489	0.444	71.6	0.487	0.438	55.2	0.488	0.445	76.6

Table 18. **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%	0.1398	0.1733	0.1773	0.1857
	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	0.1620	0.0901

variance. Table 20 reports the spatial misdirection results, split into the target-absent and target-present metrics. Table 21 reports the duration inflation results; the clean row corresponds to the unpoisoned reference model, and its ASR is the realized false-positive rate at the same delay margin. Results are stable across the three seeds overall, with the exception of the visual trigger's target-present redirection (ASR_{TP}), which varies more significantly from seed to seed (51.2 ± 13.9).

Table 19. **Case study significance tests across trigger modalities and poisoning ratios.** We report the attack success rate (ASR) and floor rate, along with McNemar's test (b, c, χ^2_1, p_{McN}) and Wilcoxon signed-rank test (z, p_{W}, r).

Trigger	ρ	**ASR**	**Floor**	b	c	χ^2_1	p_{McN}	z	p_{W}	r
Visual	2.5%	7.8%	5.5%	20	13	1.5	0.296	+0.90	0.369	+0.06
Visual	5%	8.1%	5.5%	21	13	1.9	0.229	−3.57	3.6×10^{-4}	−0.23
Visual	10%	9.4%	5.5%	26	14	3.6	0.081	+1.87	0.061	+0.12
Text	2.5%	39.4%	6.2%	113	11	83.9	1.8×10^{-22}	+11.07	1.7×10^{-28}	+0.73
Text	5%	61.9%	6.2%	176	5	161.6	1.0×10^{-45}	+13.93	4.5×10^{-44}	+0.92
Text	10%	83.1%	6.2%	237	1	234.0	1.1×10^{-69}	+14.94	1.9×10^{-50}	+0.98
Multimodal	2.5%	34.5%	5.9%	94	6	77.4	2.0×10^{-21}	+11.47	1.8×10^{-30}	+0.76
Multimodal	5%	49.2%	5.9%	141	8	118.7	1.5×10^{-32}	+12.66	1.0×10^{-36}	+0.83
Multimodal	10%	83.1%	5.9%	237	0	237.0	9.1×10^{-72}	+14.92	2.4×10^{-50}	+0.98

Table 20. **Spatial misdirection across training seeds** ($\rho = 10\%$, seeds 42/43/44). Util is the clean-sample BBox hit ratio; $\mathrm{ASR}_{\mathrm{TA}}$ and Freq. (% of target-absent images with $\mathrm{SS}_{\mathrm{atk}} > \mathrm{SS}_{\mathrm{qry}}$) are the target-absent metrics; $\mathrm{ASR}_{\mathrm{TP}}$ and ΔDist (px) are target-present. $\Delta\mathrm{Dist} = d_{\mathrm{trig}} - d_{\mathrm{clean}}$ (px; more negative = gaze pulled onto the attacker target). Last row per trigger is mean±std over the three seeds.

			Target-Absent		**Target-Present**	
Trigger	Seed	Util ↑	$\mathrm{ASR}_{\mathrm{TA}}$ ↑	Freq. ↑	$\mathrm{ASR}_{\mathrm{TP}}$ ↑	ΔDist ↓
Visual	42	0.815	63.1	62.3	67.2	−25.9
	43	0.753	59.6	58.6	42.7	−18.7
	44	0.788	61.1	57.9	43.7	−24.3
	mean	0.785±0.031	61.3±1.8	59.6±2.4	51.2±13.9	−23.0±3.8
Text	42	0.781	58.7	65.2	73.2	−25.6
	43	0.758	52.9	61.3	75.2	−31.5
	44	0.830	57.3	59.8	68.5	−28.2
	mean	0.790±0.037	56.3±3.0	62.1±2.8	72.3±3.4	−28.4±3.0
Multimodal	42	0.786	55.4	63.2	77.8	−30.4
	43	0.799	53.5	62.7	65.6	−30.2
	44	0.824	56.7	63.2	74.5	−28.6
	mean	0.803±0.019	55.2±1.6	63.0±0.3	72.6±6.3	−29.7±1.0

Table 21. **Duration inflation across training seeds** ($\rho = 10\%$, seeds 42/43/44). Utility is the clean-sample temporal scanpath similarity SS_t; ASR is the percentage of triggered inputs whose induced delay exceeds the clean margin δ=11.5 ms over n=612 test pairs. Last row per trigger is mean±std over the three seeds. The clean row is the unpoisoned reference model, whose ASR is the realized false-positive rate at the same δ. Seed-to-seed spread within a trigger exceeds the spread between triggers, so three seeds are not enough to rank the triggers.

Trigger	Seed	Utility (SS_t) ↑	ASR ↑
Clean	–	0.451	5.5
Visual	42	0.441	87.1
	43	0.434	95.3
	44	0.438	94.8
	mean	0.437±0.004	92.4±4.6
Text	42	0.439	95.1
	43	0.439	90.7
	44	0.435	91.0
	mean	0.438±0.003	92.3±2.5
Multimodal	42	0.436	89.5
	43	0.440	92.6
	44	0.430	90.5
	mean	0.435±0.005	90.9±1.6

References

[1] 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. 5

[2] 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. 7

[3] 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

[4] 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. 5, 6

[5] 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. 5

[6] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *RAID*, 2018. 4

[7] 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

[8] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems*, 35:10546–10559, 2022. 4

[9] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 4

[10] 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. 5