Reference-beam attacks against twin-field quantum key distribution using optical injection locking

Sergio Juárez,[1,2,*,†] Alessandro Marcomini,[3,2,4,*,‡] Mikhail Petrov,[3,2,4] Robert I. Woodward,[1] Toby J. Dowling,[1] R. Mark Stevenson,[1] Marcos Curty,[3,2,4] and Davide Rusca[3,2,4]

[1]*Toshiba Europe Limited, 208 Cambridge Science Park, Cambridge CB4 0GZ, United Kingdom*

[2]*Escuela de Ingeniería de Telecomunicación, Department of Signal Theory and Communications, University of Vigo, Vigo E-36310, Spain*

[3]*Vigo Quantum Communication Center, University of Vigo, Vigo E-36310, Spain*

[4]*AtlanTTic Research Center, University of Vigo, E-36310, Spain*

(Received 1 September 2025; accepted 6 February 2026; published 12 March 2026)

Twin-field quantum key distribution (TF-QKD) has become a leading protocol to bring quantum communications to the national scale. The protocol requires the establishment of a shared phase and frequency reference between distant parties, which is commonly achieved by using an external reference laser in an optical injection locking (OIL) architecture. In this work, we analyze the side channels in OIL-based TF-QKD that may arise from adversarial manipulation of the various degrees of freedom of this untrusted reference beam. We experimentally demonstrate two realistic attack scenarios: fast intensity modulation of the reference laser, and additional signals embedded in the reference light exploiting wavelengths undetectable by conventional monitoring techniques. These attacks can allow a potential eavesdropper to deterministically increase the mean photon number of the sources, or circumvent the decoy-state technique, respectively. To counter these vulnerabilities, we propose practical and highly effective countermeasures that reinforce the security of TF-QKD systems without significant additional complexity or performance degradation.

DOI: 10.1103/71m5-3c5n

I. INTRODUCTION

Quantum key distribution (QKD) enables two legitimate parties, Alice and Bob, to share a symmetric encryption key that can be used to achieve information-theoretically secure communications. In particular, it exploits the laws of quantum mechanics to ensure that any eavesdropping attempt by a malicious external party (Eve) cannot go undetected [1–4].

To date, QKD has matured into a well-established technology whose application has been demonstrated on the level of intercity networks [5–7]. Next-generation protocols such as twin-field QKD (TF-QKD), which utilize single-photon interference, offer significantly improved resilience against channel losses [8–11], effectively doubling the achievable communication distances [12,13], reaching up to 1000 km point-to-point [14–18]. These properties make TF-QKD particularly attractive for the realization of nationwide quantum networks [19–23].

Nevertheless, the broader adoption of QKD still faces some challenges, particularly the need to guarantee the security of practical implementations, since real devices inevitably deviate from the idealized assumptions. These deviations must be accounted for, either through improved security proofs or through enhancements in physical hardware, a domain collectively known as implementation security [24,25]. Unaddressed device imperfections can lead to unintended side channels, resulting in undetected information leakage, which may compromise the overall system security [26–28].

In this regard, TF-QKD possesses a crucial advantage, as it belongs to the measurement-device-independent (MDI) class of protocols. MDI protocols remove all security assumptions from the receiver, ensuring security even if Eve completely controls the measurement apparatus [29]. Consequently, implementation security efforts can be entirely focused on the transmitter side.

Practical implementations of TF-QKD require a shared phase and frequency reference between Alice's and Bob's encoders. Typically, this reference is provided by the untrusted central measurement node (Charlie), which distributes a classical reference beam through a dedicated service channel. The received beam stabilizes the laser sources via optical phase-locked loops (OPLLs) or, more commonly, optical injection locking (OIL) [30–32]. OIL is particularly favored due to its inherent stability, practicality, and lack of necessity for active feedback control [19,33]. However, permitting an external beam, potentially altered by Eve en route, to directly enter the encoders raises critical concerns regarding implementation security [27,34,35].

In this work, we provide a thorough analysis of implementation security specifically tailored to the OIL architecture used in TF-QKD. We investigate potential attack scenarios

*These authors contributed equally to this work.
†Contact author: sergio.juarez@toshiba.eu
‡Contact author: amarcomini@vqcc.uvigo.es

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

enabled by manipulation of the reference laser, examining multiple optical degrees of freedom and revealing two potential side channels. Additionally, we also provide practical countermeasures for them.

The first identified side channel exploits a fast intensity modulation (FIM) of the reference laser, performed on the same timescale as the pulse encoding or faster, which temporarily increases the emitted photon number without detection by conventional monitoring.

The second side channel exploits back-reflections from the laser diode (LD) in the OIL configuration, and the limited spectral response of monitoring devices. The separate channel that goes into the encoders allows Eve to embed an additional signal at a Trojan wavelength in the reference light (TWIRL). When this undetected light reaches the encoder, it could possibly leak the complete information about the intensity and phase settings used in the quantum signals. Note that this channel cannot be blocked without defeating its purpose.

Our analysis underscores the necessity for careful spectral filtering of the injected reference beam and highlights the importance of employing monitoring systems ("watchdogs") that track the timing and integrity of the incoming signals with high temporal resolution. Implementing these countermeasures effectively closes these side channels, significantly strengthening the security of OIL TF-QKD implementations.

The remainder of this paper is structured as follows: first, we provide a detailed examination of potential attack vectors originating from Eve's manipulation of Charlie's reference signal, identifying FIM and out-of-band TWIRL signals as realistic threats. Next, we experimentally demonstrate the feasibility of these attacks and evaluate suitable countermeasures. Finally, we present our conclusions and broader implications of our findings for the practical security of TF-QKD.

II. POTENTIAL SIDE CHANNELS IN OIL TF-QKD

In Fig. 1(a), we present a simplified schematic of a typical TF-QKD system consisting of Alice, Bob, and Charlie. The central node (Charlie) provides a coherent optical reference to the transmitters (Alice and Bob) through the service channel. Alice and Bob individually prepare quantum states, which are subsequently sent back to Charlie via the quantum channel for single-photon interference and measurement. Both channels can potentially be manipulated by Eve, and, in the worst case, Charlie himself could be entirely under Eve's control. Figure 1(b) details the most important components of an encoder and receiver within an OIL-based TF-QKD architecture. The optical reference received from Charlie first passes through a polarization controller (PC) and a polarizing beam splitter (PBS), ensuring single-polarization injection to optimize the injection-locking efficiency. The optical reference goes then through a circulator and is injected into the encoder's laser, locking its emission with that of Charlie's laser. This results in a shared optical frequency and a fixed phase offset, which enables coherent operation across all sources. Subsequently, the locked signal passes through the encoder, which uses intensity and phase modulators to craft the train of pulses through pulse carving and manual (active) phase randomization. Finally, the pulses get attenuated to the single-photon level using a variable optical attenuator (VOA) and are

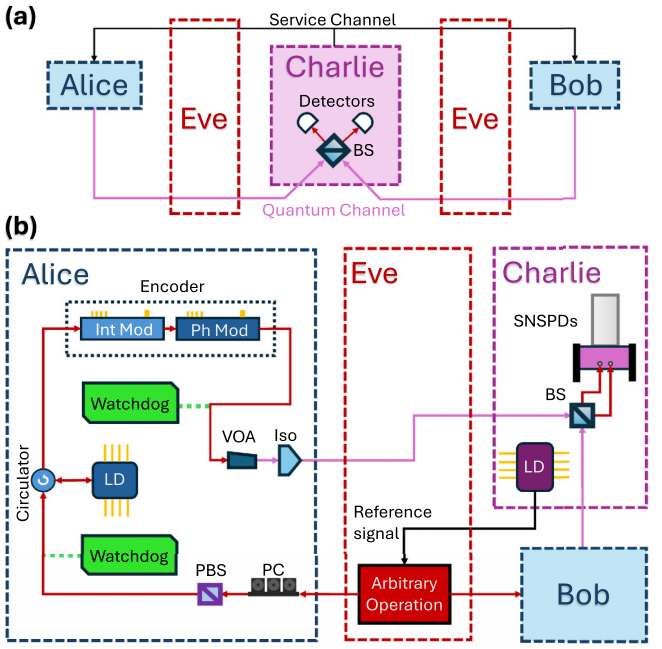


FIG. 1. Schematics of a basic TF-QKD system and encoder architecture based on OIL. (a) Overview of a typical TF-QKD setup where Charlie distributes a coherent reference via the service channel. Alice and Bob encode quantum signals and send them back through the quantum channel for interference at Charlie's beam splitter (BS) and measurement. Eve can manipulate both channels and, in the most extreme case, take full control of Charlie. (b) Key components of an OIL-based encoder. The reference passes through a PC and PBS for single-polarization injection, then locks the local LD via a circulator. The output is encoded using an intensity modulator (Int Mod) and phase modulator (Ph Mod), attenuated to the single-photon level with a VOA, and sent to Charlie through an isolator (Iso). Interference and detection are performed at Charlie using a BS and single-photon detectors such as SNSPDs. The watchdogs guarantee the integrity of these signals. Bob's setup is identical to Alice's.

transmitted back to Charlie after passing through an isolator that prevents attacks from the quantum channel against the transmitter.

This architecture simplifies frequency and phase stabilization, ensuring coherent interference at Charlie's measurement node. However, adversaries might attempt to exploit any of the various degrees of freedom of the injected reference signal, such as polarization, phase, intensity, or wavelength (frequency), to gain information through side-channel attacks.

Polarization-based attacks are mitigated by the PBS at the input of the security perimeter, which transmits a single polarization component while suppressing the orthogonal one with a typical extinction ratio of 30 dB. As a result, polarization attacks relying on rapid fluctuations of the polarization of the reference beam lose one projection, while the remaining accepted component only fluctuates in intensity. Consequently, polarization attacks become equivalent to intensity attacks. Phase manipulation attacks are ineffective due to the active phase randomization in Alice's and Bob's encoders after the OIL stage, and any excessive phase fluctuations are detectable as they will increase the quantum bit error rate (QBER), resulting at most in a denial-of-service scenario.

Attacks targeting the reference laser's intensity could allow Eve to induce transient increases in the output power of Alice's and Bob's lasers, implying a threat to the single-photon level encoding of QKD. For instance, increasing the intensity of the coherent state $\langle\mu\rangle = 0.4 \rightarrow 0.6$ doubles the multiphoton emission probability. A FIM attack that goes undetected would thus enable Eve to circumvent the security provided by the no-cloning theorem and the decoy state method [9,36–38].

Spectral attacks encompass two distinct scenarios: manipulating the frequency of the injected reference itself, and injecting additional optical signals at wavelengths far removed from the intended reference wavelength. While the former approach inevitably worsens the interference at the central node, thus enlarging the QBER, the latter exploits vulnerabilities arising from wavelength-dependent responses of the detectors and variations in attenuation across optical components within the system. Importantly for TF-QKD systems, using a service channel to introduce a reference into the system creates an entry point for TWIRL signals that cannot be fully closed.

To mitigate these potential attacks, we recommend the use of monitoring detectors ("watchdogs") strategically placed within the encoder setup, as illustrated in Fig. 1(b). A first watchdog placed before optical injection ensures the integrity of the incoming reference signal, while a second watchdog positioned immediately after the encoder monitors the output pulse train before attenuation to the single-photon level, enabling classical-level signal processing. These watchdogs should continuously monitor power, timing, and spectral characteristics, triggering alerts upon detecting anomalies indicative of hardware malfunctions or potential eavesdropping attempts.

However, the design of these watchdogs requires careful consideration, as Eve could exploit their intrinsic limitations to evade detection [37]. For instance, slow photodiodes (PDs) that integrate optical power over relatively long time intervals, compared to the encoding rate, might be sufficient to treat slow intensity fluctuations, but they will fail to detect FIM. Eve could maintain a constant average optical power, lowering and increasing the power within time windows shorter than the integration period, thus avoiding detection. Similarly, PDs with responsivity optimized for specific wavelengths (typically aligned with the intended communication wavelength) can be vulnerable to wavelength-dependent Trojan-horse attacks (THAs). Eve could impersonate Charlie by injecting anomalous signals at wavelengths outside the PD's detection range. In such scenarios, the watchdog would register only minimal increases in optical power, despite the presence of a high-intensity TWIRL signal. This extrinsic signal could then propagate undetected through Alice's and Bob's setups, surviving the wavelength-dependent attenuation of critical optical components such as VOAs, intensity modulators, and phase modulators. Consequently, the TWIRL signal could reach deep into the encoders, extracting sensitive information about the encoded quantum states (including intensity setting, basis choice, and even bit choice) and leaking it back to Eve.

To accurately determine the required performance characteristics of these watchdogs and enhance their resilience, we experimentally recreated these attack scenarios under controlled conditions. Using detectors with different temporal and spectral responses, we evaluated the effectiveness of these attacks and established robust design criteria for the watchdogs, as presented in the following section.

III. EXPERIMENTAL DEMONSTRATION AND COUNTERMEASURES

A. Fast intensity modulation attack

In Fig. 2 we present the setup used to investigate the FIM attack, where Eve modulates only the intensity of the reference laser. Such signal is routed into Alice's (equivalently, Bob's) station, passing through a PC and a PBS to ensure a single polarization, effectively converting polarization-based attacks into intensity-based ones. A narrowband spectral filter is applied to reroute unwanted spectral components (see below) and isolate the expected reference wavelength, preventing injection of extrinsic light. The spectrally filtered signal is then used to injection-lock Alice's laser, synchronizing its phase and frequency with the modulated reference. Importantly, all lasers operate in continuous wave (CW) mode, and the encoder and the rest of the components shown in Fig. 1(b) are omitted in the setup shown in Fig. 2. This is because the experiment focuses specifically on assessing how FIM of the reference signal affects the OIL process, potentially compromising transmitter security.

Superconducting nanowire single-photon detectors (SNSPDs) equipped with a time tagger, together with an optical spectrum analyzer (OSA), are used to characterize the properties of the modulated reference signal prior to the locking of Alice's laser, and also to study the effects that the attack has on the effective photon numbers and the spectra of the locked laser. Using slow, integrating power meters, we monitor the power levels of both the unfiltered optical signal and, in conjunction with the filters, any contributions from extrinsic or unexpected wavelengths, measured both before and, more importantly, after the locking stage.

The first option we investigate for a potential watchdog consists of two GHz-bandwidth fast PDs monitoring the high-speed behavior of the system: the injection PD tracks the temporal profile of the signal arriving at the encoder's laser via the circulator, while the locking PD observes the output of the laser after the locking process. To assess both the threat posed by FIM attacks and the effectiveness of these fast PD as countermeasures, we experimentally characterized an injection-locked laser under attack conditions. Specifically, to resolve a 1 GHz modulation of the reference laser and its impact on the encoding laser, we employed fast PDs with bandwidths of 10 GHz and 40 GHz, respectively.

The detection results of an injection locked laser under a FIM attack [Fig. 3(a)] are displayed in Figs. 3(b) and 3(c). The SNSPDs data in Fig. 3(b) indicates that Eve's modulation can produce transient increases in the mean photon number by up to 51% during the modulated time intervals. This demonstrates that the FIM attack is a credible threat, as it enables a controlled increase in the photon number emitted by Alice or Bob. Conventional power meters register no modulation when integrating in their fastest configuration of 25 μs. The results of these measurements are in the Supplemental Material [39]. Although we observe power meters to be oblivious to these modulations, we note that even if they could detect sustained

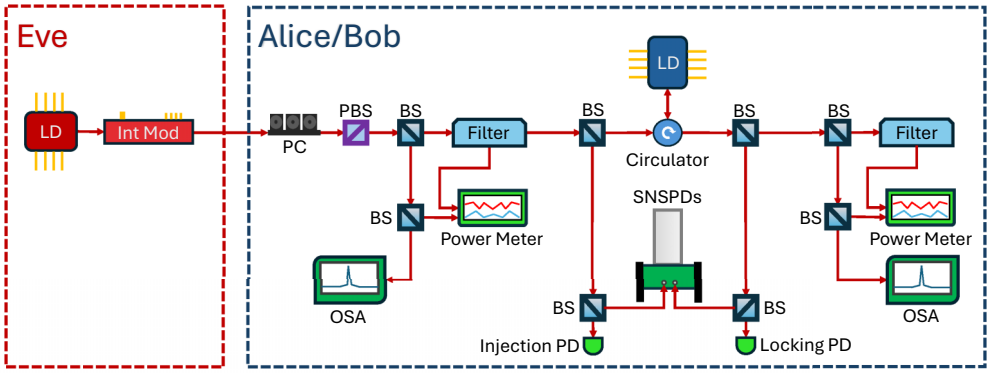


FIG. 2. Experimental setup for studying the FIM attack. Here, Eve is assumed to fully control the intensity of the reference laser, implemented using an intensity modulator (Int Mod). The modulated reference signal enters Alice (or Bob), passing through a PC and a PBS, then undergoes spectral filtering to isolate the agreed reference from extrinsic wavelengths before injection locking the LD. The behavior of the injected and locked signals is characterized by SNSPDs and an OSA. The signals are also monitored by two PDs with GHz detection bandwidths, one to observe the signal injected into the encoder's laser (Injection PD) and another to monitor the laser's output after locking (Locking PD), and a power meter that checks filtered and unfiltered signals before and after the locking. After the second filter, the signal would normally pass through the encoder, be attenuated, and then sent to Charlie, as in Fig. 1(b).

increases or decreases in average power, Eve could strategically time the modulated pulses to dilute the fluctuations, making them indistinguishable from detector noise. Therefore, our analysis shows that slow-integrating power meters alone are insufficient against this type of attack.

The effectiveness of this attack, and the failure of slow-integrating power meters to detect it, can be understood from fundamental physical constraints on Eve's strategy. First, to affect the photon number of an encoded pulse, Eve's modulation must occur within the temporal window of that pulse. This confines any effective attack to timescales on the order of the pulse duration, or shorter. Furthermore, to evade detection by slow-integrating power meters, Eve must maintain a constant average optical power. This requirement forces her to employ symmetric modulation patterns such as "Up-to-Down" or "Down-to-Up" [see Fig. 3(a)], where any transient increase in power is compensated by a corresponding decrease. As a consequence, the total excess energy that Eve can redistribute into any given pulse is limited by the product of the nominal reference power and the pulse duration. An extensive overview of the pulse designing phase, together with additional modulation scenarios, can be found in Appendix. Eve is also constrained by the physics of OIL. In fact, Eve does not have direct control over Alice's laser output: she can

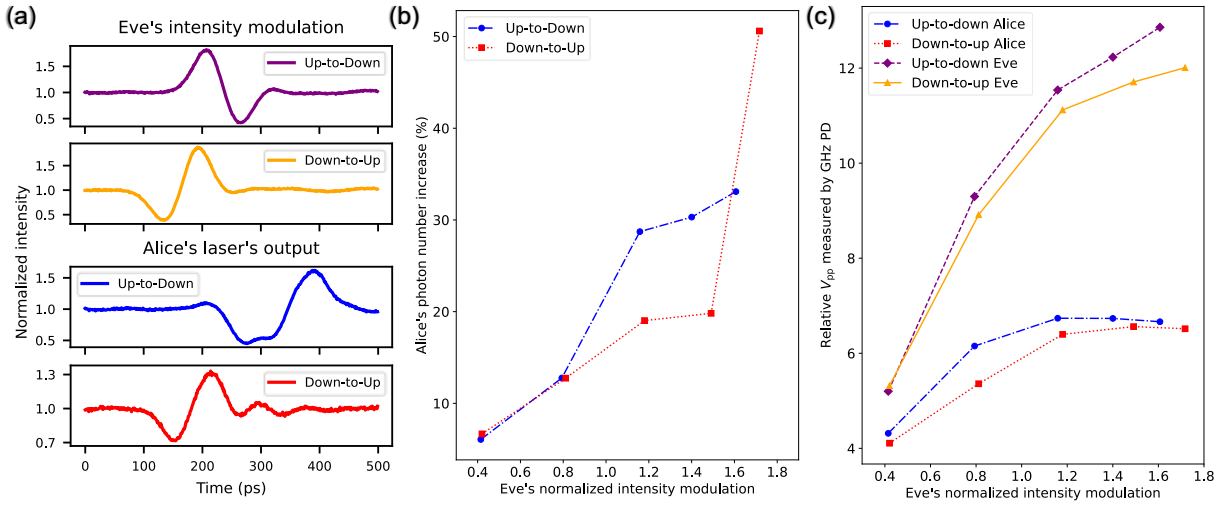


FIG. 3. Detection performance of watchdog detectors under FIM attacks. We study the attack under various amplitude modulations: "Eve's normalized intensity modulation" in all subfigures (and in Fig. 6) refers to the difference in power between the minima and maxima of the modulated laser signal (peak-to-peak), relative to the baseline power of the unmodulated signal. (a) Examples of the two primary intensity modulation patterns tested: "Up-to-Down" (power increased, then reduced) and "Down-to-Up" (power reduced, then increased), both maintaining constant average intensity. (b) SNSPD measurements showing a transient increase in mean photon number (up to 51%) during the modulation intervals. Slow-integrating power meters failed to detect these modulations. The abrupt increase at the final point of the "Down-to-Up" configuration is due to the merging between the main pulse and the trailing ringing. (c) Fast-PD output reliably identifying all modulation types tested, enabling immediate attack detection and limiting adversarial interference to a denial-of-service scenario.

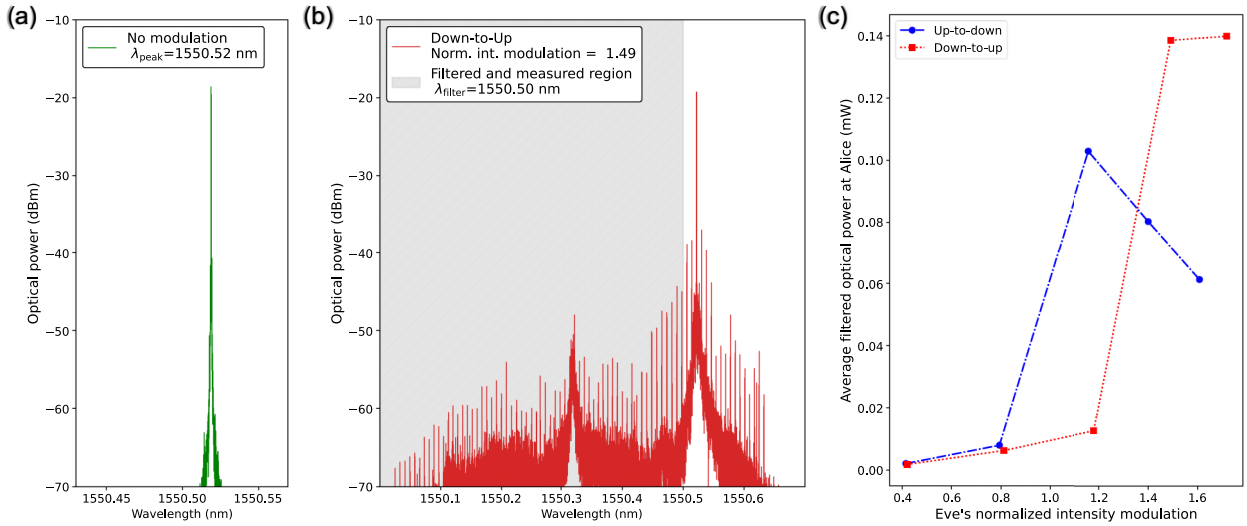


FIG. 4. Spectral detection of FIM attacks. (a) Reference optical spectrum of Alice's LD without modulation. (b) Spectrum under active FIM at 1 GHz repetition rate, showing prominent sidebands around the main wavelength. (c) Power detected in the filtered sidebands region using a slow (25 µs integration) power meter. The observed increase in power clearly indicates modulation-induced spectral broadening, providing a simple yet effective secondary detection method. The decay observed towards the end results from variations in the sidebands moving outside the tunable filter's passband; these features are also visible on the right side of the spectrum.

only modulate the injected reference signal, and the laser's output response is mediated by the injection locking dynamics. These dynamics (including the finite locking bandwidth, the cavity photon lifetime, and amplitude-phase coupling in the gain medium) impose a minimum response time that prevents arbitrarily fast modulations of the locked laser output, regardless of how rapidly Eve modulates her input.

Taking into account these constraints on the attack, a fast PD with bandwidth matching (or exceeding) the system clock rate and with a rise time comparable to (or shorter than) the pulse duration can resolve intensity variations within each pulse period, and will therefore detect Eve's manipulation. Crucially, if Eve attempts to evade detection by compressing her modulation into shorter timescales, she faces a fundamental tradeoff: shorter modulation windows proportionally reduce the energy she can modulate per pulse. In other words, the same strategy that might help Eve avoid detection simultaneously renders her attack ineffective.

As reported in Fig. 3(c), the fast PDs reliably detect anomalies for every modulation tested, promptly identifying the attack, enabling countermeasures, and limiting Eve's impact to a denial-of-service scenario since once the attack is detected the protocol must be interrupted.

Furthermore, owing to amplitude-phase coupling in the laser's gain medium, oscillations in the injected power convert partly into instantaneous frequency shifts, resulting into both intensity modulations and spectral sidebands [32]. This spectral distortion in Alice's local laser output provides another signature of malicious intervention, which can be detected as a countermeasure. To this end, the combined use of the slow-integrating power meter alongside spectral filters to detect unexpected wavelength components results in a secondary detection venue against FIM attacks.

The measurement results of the power meter aided by the spectral filters are shown in Fig. 4. Comparing an unmodulated spectral baseline [Fig. 4(a)] with an active modulation scenario [Fig. 4(b)], clear sidebands emerge. A narrow-band spectral filter directs these otherwise absent spectral sidebands into a power meter, providing a straightforward detection method even with slow integration times. The results of Fig. 4(c) were taken with the 25 µs integration of the power meter, but are observable with integration times up to 100 µs. This approach enables reliable identification of FIM attacks by monitoring unexpected spectral shifts, enabling Alice and Bob to abort the QKD session for protection, since any increase in power in these wavelengths would be created by an attack on the encoding laser.

For practical implementations, we recommend a calibration procedure in which the injected reference is modulated at various depths while monitoring the out-of-band power through a narrow-band filter positioned just outside the unperturbed laser's spectral range. This establishes, for the specific laser and OIL configuration, the minimum modulation depth producing detectable spectral distortion, ensuring the watchdog threshold is appropriately set.

Crucially, underestimating the actual optical power emitted by Alice and Bob, even if still under the single-photon regime, can lead to an overestimation of the achievable secret key rate (SKR). To illustrate this, we consider the "Sending-or-Not-Sending" TF-QKD protocol, and compute the asymptotic SKR in the presence of the attack (see Ref. [9] for the protocol description and security proof). The results of this analysis are presented in Fig. 5, with details about the simulations provided in Appendix. Plots in Fig. 5 demonstrate that, even in the asymptotic limit, there is a noticeable discrepancy in the SKR estimation under attack conditions. Such discrepancies are expected to become more pronounced in finite-key scenarios, since one has to heavily rely on measurements in the control basis to provide meaningful estimates of the phase error rate, and since this procedure is based on studying the interference of the weak coherent pulses sent by Alice and Bob, the FIM-induced amplification of such pulses will result in

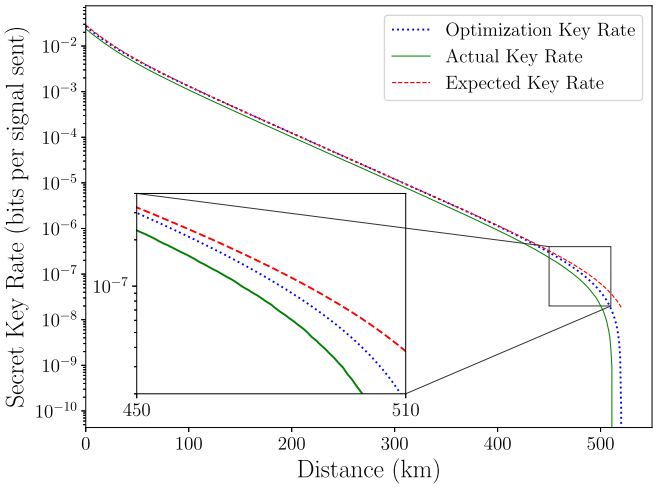


FIG. 5. FIM attacks lead to an erroneous estimation of the SKR. To run their protocol, Alice and Bob must first optimize some parameters, based on their system characteristics and channel conditions. This optimization yields a theoretical estimate for the expected SKR (blue dotted curve). However, under Eve's FIM attack the signal states they actually prepare have enlarged intensity. Here we set the enhancement factor to +51%, corresponding to the maximum intensity increase experimentally measured in Fig. 3(b). If Alice and Bob are aware of the intensity increment, they can correctly apply the security analysis, obtaining a lower bound on the SKR that accurately reflects the attack (green solid curve). Nevertheless, if they were oblivious to the attack and employed the expected intensity value in the SKR calculation together with the actual statistics observed, they would retrieve an erroneous SKR, which is higher that the actual one provided by the proof (red dashed line), thus overestimating it. Further details and simulation parameters are provided in Appendix.

out-of-ordinary estimates of the protocol's parameters, potentially severely impacting performance.

While a full protocol-specific security analysis accounting for finite-size effects is beyond the scope of this work, we note that in extreme cases with small data block sizes, the induced discrepancy between the expected and the actual SKR could even lead to the former surpassing the theoretical upper bound on the achievable key rate, yielding a completely insecure [12,38] estimate. We cannot evaluate such claim for the case at hand, as no explicit upper bound on the SKR for TF-QKD is available, to be best of our knowledge. Although this means that no definitive insecurity claims can be done at present time, it has to be anticipated that a carefully designed attack could increase the emitted intensity beyond the values in Fig. 3(b), and the upper bound on the SKR will thus eventually be violated. Therefore, accurately bounding the emitted states' intensity is critical to ensure practical security.

B. Trojan wavelength in the reference light attack

We now move on to characterize the TWIRL attack in the context of OIL TF-QKD, where we experimentally investigate the spectral response of the laser cavity itself. Critically, any Trojan-wavelength light, which bypasses the watchdogs and gets reflected by the laser cavity will be injected into the encoder and undergo the same encoding steps, potentially leaking full information about the intensity and phase settings if not effectively filtered. It should be noted that this component arrangement is unique to TF-QKD, as the reference signal reaches the LD via the circulator connected to the service channel, rather than from the quantum channel.

Previous studies have documented the spectral behavior of common encoder components such as attenuators and modulators [40]. However, the response of a laser cavity under illumination by wavelengths far outside its typical operating band has not yet been characterized, particularly in the context of OIL architectures. As this is the fundamental missing piece to properly compute the total attenuation along the full optical path experienced by light at every wavelength, our setup [depicted in Fig. 6(a)] isolates the laser cavity using only a fiber-optic circulator and a distributed feedback (DFB) laser, deliberately omitting other encoder components in order to specifically analyze cavity interactions at out-of-band wavelengths. In doing so, we aim to reveal any previously unidentified spectral dependencies or unexpected vulnerabilities inherent in OIL-based TF-QKD architectures.

To explore a broad spectral region, we use a supercontinuum wideband fiber laser covering a wavelength range of 500–2200 nm. While all encoder components are optimized for the telecom C band centered around 1550 nm, we intentionally illuminate the DFB laser with wavelengths far from its nominal injection locking region. Importantly, the laser under test does not have an isolator and is not locking to this out-of-band light. Reflected signals, largely bypassing the cavity due to off-resonance conditions, are collected via the circulator and measured using an OSA. This arrangement allows us to directly measure and characterize how the laser cavity structure affects signals at wavelengths significantly outside its normal operational range, identifying spectral regions where unexpectedly high levels of reflected power could aid Eve's TWIRL attack.

It is worth noting that while our experimental characterization was performed using a DFB laser (the most commonly used laser type within QKD systems), the TWIRL attack is expected to apply broadly to any laser used in OIL architectures. This generality arises because OIL inherently requires an optical pathway into the laser cavity, and this same pathway is accessible to light at TWIRL wavelengths. Any laser cavity will exhibit some degree of reflectivity at out-of-band wavelengths via interaction with cavity interfaces or facets, particularly since the gain medium provides neither absorption nor amplification far from the operating wavelength. The quantitative reflectivity spectrum will depend on the specific laser design, and therefore a thorough security assessment should include characterization of the laser employed in each implementation.

Figure 6(b) shows the spectral losses observed by comparing spectra from the wideband source alone to those from three different experimental configurations: circulator only (no LD connected), circulator with LD connected but powered off, and circulator with LD actively lasing. Conversely, Fig. 6(c) highlights the effective transmissivity gained through the presence of the LD cavity in the OIL configuration, calculated by comparing configurations with the LD connected and disconnected (LD off only). This approach reveals the ideal wavelengths to use for this attack. That is, those with positive transmissivity gains that are undetectable by typical

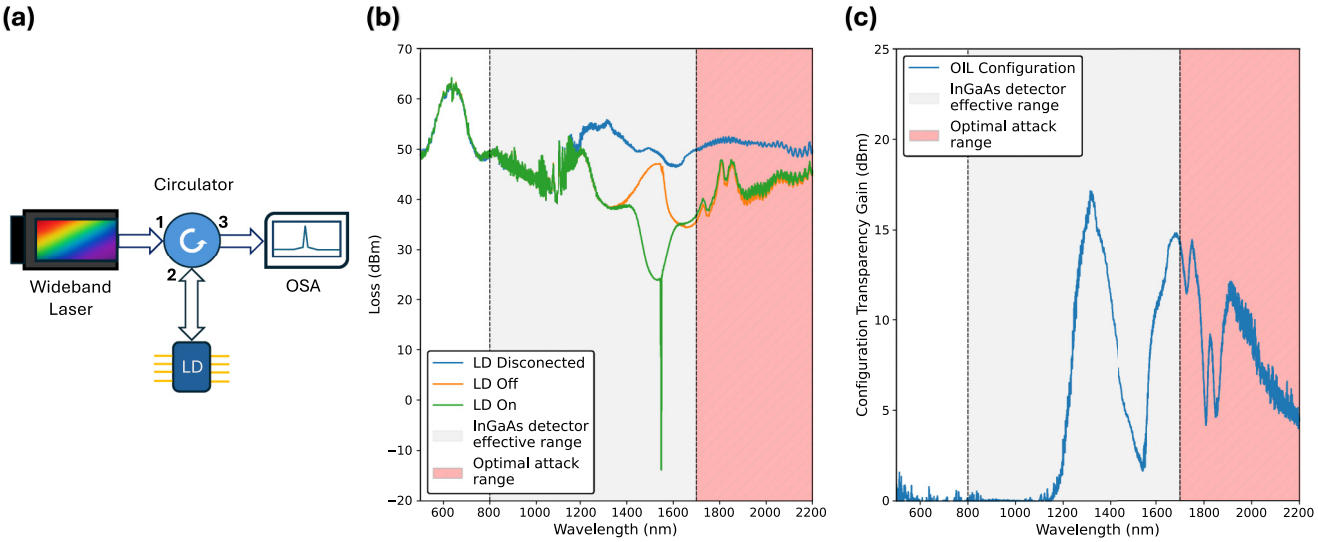


FIG. 6. Spectral characterization of laser cavity reflections relevant to TWIRL attacks. (a) Experimental setup for studying the TWIRL attack. A wideband laser and an OSA are employed to characterize the spectral response of the LD over a broad wavelength range from 500 nm to 2200 nm. (b) Measured spectral losses for different experimental configurations of the encoder setup. Losses were determined by subtracting the spectra obtained with the wideband laser alone from those measured when the circulator and LD were connected. Three configurations are compared: circulator alone (no LD connected), circulator with LD connected but powered off, and circulator with LD connected and powered on. (c) Effective transparency gained in the whole circulator-DFB LD configuration, computed as the difference between the losses with the LD disconnected and the LD connected but off. Results show significant deviation of the expected transparency due to reflections when the cavity is connected, and allow to determine the optimal wavelengths for the attack.

watchdogs (such as InGaAs PDs, which are optimized for ∼1550 nm and sharply lose responsivity beyond 1700 nm [41]), thus allowing the injected signals to bypass the cavity unnoticed.

The results in Fig. 6(c) clearly indicate significant cavity-induced reflections within 1200 nm to 2200 nm. Eve could exploit this broad spectral range to inject a high-power optical signal, with a wavelength significantly above what an InGaAs detector would detect. Under these conditions, the detected optical power would substantially underestimate the actual injected power, allowing Eve's TWIRL signal to propagate through the encoder at classical power levels, despite the typical attenuation mechanisms implemented within quantum encoding setups.

These results illustrate how the TWIRL attack must be carefully taken into account for the security of any OIL-based TF-QKD system, as it allows us to completely learn the encoding settings when overlooked. Fortunately, to prevent this attack one only requires an adequate amount of spectral filtering with high enough attenuation at these external wavelengths, thus ensuring that for optical frequencies outside of the locking range, the intensity of the light after the encoder is sufficiently weak to guarantee security.

In order to compute the required amount of attenuation one must understand what is the maximum amount of power that Eve could inject into Alice and Bob's devices, as well as what is the maximum acceptable output power of the encoders, according to security proofs. For the former is typically considered, as a worst-case scenario, the so-called laser-induced damage threshold (LIDT). For light at telecom wavelengths, Ref. [35] reports a theoretical estimate of the LIDT value of 55 kW, while in a recent certification guideline Ref. [27] considers as a more reasonable conservative assumption that the maximum possible insertion power for a realistic Eve is 100 W. We remark that this power threshold only applies to fiber, and does not take into consideration possible damage to connectors and other components, meaning that in practice it is probably still overconservative.

Crucially, the value of the LIDT strongly depends on the wavelength [42]. Therefore, by assuming a LIDT of 100 W at $\lambda_0 = 1550$ nm, the LIDT at an arbitrary wavelength λ can be found by following the square-root dependence as

$$\mathrm{LIDT}(\lambda) = \sqrt{\frac{\lambda}{\lambda_0}} \times 100\ \mathrm{W}. \tag{1}$$

Note that Eq. (1) suggests that the maximum allowed input power for Eve is larger for longer wavelengths, yielding approximately 105 W at 1700 nm. Within the encoder, the VOA typically provides 70–75 dB of attenuation at the operating wavelength to reach single-photon levels; however, this attenuation is wavelength dependent, and components optimized for 1550 nm may provide substantially less isolation at out-of-band TWIRL signals [40]. When considering the worst case scenario, it is standard to assume that Eve can bypass channel losses entirely, placing herself right outside Alice's security perimeter. However, for an attacker operating from Charlie's position, the attack signal would face additional fiber losses at out-of-band wavelengths due to infrared absorption (approximately 0.75 dB/km at 1700 nm versus ∼0.2 dB/km at 1550 nm [43]). Noting the differences in loss across wavelengths, it is more conservative to provide suppression of unexpected out-of-band wavelengths inside the encoder, assuming the LIDT threshold.

The TWIRL attack is conceptually equivalent to the standard THA coming from the quantum channel. Therefore, to produce a quantitative and theoretically secure upper bound for the maximum tolerable output power a comprehensive analysis against these, specific to TF-QKD, is required. Unfortunately, to date such analysis has yet to be established, meaning that no formal estimation of the induced information leakage is currently available for this scheme. One can get a rough estimate of the required isolation from security proofs for other QKD schemes, under the assumption that the maximum allowed output power of a TF-QKD transmitter in the Trojan mode is comparable to that of prepare-and-measure QKD. For these, an overall input-to-output optical isolation of approximately 200 dB is typically considered necessary to approach the performance of an ideal, leakage-free scenario [35,44–49]. Fortunately, such levels of attenuation can be achieved using off-the-shelf spectral filters. When cascaded, multiple units can provide the required out-of-band suppression with less than 10 dB of total in-band loss, while completely suppressing TWIRL signals [40]. In any case, we remark that to properly close this security loophole for a given system, one should consider the maximum tolerable Trojan light intensity established by a dedicated security proof, and design the attenuation scheme accordingly.

IV. CONCLUSIONS

In this work, we have performed a detailed analysis of all possible degrees of freedom of the reference laser that could pose a security threat in OIL-based TF-QKD systems. We identified two viable attack vectors: FIM of the injected reference, and Trojan signals at wavelengths well outside the optimal detection range of usual watchdogs, equipped with sufficient power to pass through the system despite existing attenuation.

To counter external FIM of the generated pulse train, we recommend the use of a fast PD as a watchdog, ideally as fast as the encoding itself, in order to take advantage of the fact that the reference signal is still at classical power levels before attenuation to the single-photon regime. Additionally, we propose placing a narrow-band filter (say, e.g., 0.05 nm wide) centered on the reference wavelength, with all sidebands redirected to a dedicated detector: as we have shown, this allows for detection of spectral fluctuations in the output of the encoder laser that are generated by the manipulation of the OIL.

Moreover, our results reveal that out-of-band reflections from the OIL cavity create a clear pathway for high-power Trojan signals. Fortunately, this vulnerability can be effectively neutralized with straightforward measures. By limiting incoming light exclusively to the spectral range of the watchdog PD, simple and commercially available optical filters can readily suppress any unintended wavelengths. These devices introduce minimal additional optical loss, typically in the order of a few tenths of a decibel, and can be seamlessly integrated into existing setups without substantial redesign. This, coupled with a periodic spectral verification and monitoring of Alice and Bob's lasers, comprehensively addresses the identified vulnerabilities, significantly reinforcing the practical security of OIL-based TF-QKD systems without compromising their performance.

Although the proposed countermeasures will allow us to close the identified security loopholes, we remark that their proper implementation is conditional on an accurate analysis of the system specifics and the application of an adequate security proof, establishing concrete values for the amount of required attenuation and filtering. To the best of our knowledge, to date no such proof exists, and while its derivation is beyond the scope of this paper, our results motivate the inclusion of information leakage induced by reference-beam attacks into future theoretical analyses of TF-QKD.

More broadly, our analysis emphasizes the importance of restricting Eve's control over the degrees of freedom of the injected optical signals, such as frequency, power, temporal shape, and polarization, to ensure robust security of quantum communication protocols. A rigorous and comprehensive restriction not only addresses currently identified attack vectors but also provides resilience against future, yet-undiscovered, attack strategies. Maintaining tight control over all optical parameters entering the encoder thus forms a foundational element for securing practical quantum communication systems in real-world deployments.

ACKNOWLEDGMENTS

The authors thank S. Morrissey, F. Grünenfelder, O. Crampton, A. Brzosko, Y. S. Lo, G. Shooter, and P. R. Smith, for insightful discussions. We acknowledge support from the European Union's Horizon Europe Framework Programme under the Marie Skłodowska Curie Grant No. 101072637, Project Quantum-Safe Internet (QSI). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. We also acknowledge support from the Galician Regional Government (consolidation of research units: atlanTTic), the Spanish Ministry of Economy and Competitiveness (MINECO), the Fondo Europeo de Desarrollo Regional (FEDER) through the Grant No. PID2024-162270OB-I00, MICIN with funding from the European Union NextGenerationEU (PRTR-C17.I1) and the Galician Regional Government with own funding through the "Planes Complementarios de I+D+I con las Comunidades Autonomas" in Quantum Communication, the "Hub Nacional de Excelencia en Comunicaciones Cuanticas" funded by the Spanish Ministry for Digital Transformation and the Public Service and the European Union NextGenerationEU, the European Union's Horizon Europe Framework Programme under the project "Quantum Secure Networks Partnership" (QSNP, Grant Agreement No. 101114043) and the European Union via the European Health and Digital Executive Agency (HADEA) under the Project QuTechSpace (Grant No. 101135225).

S.J., A.M., R.I.W., and D.R. identified the academic motivation for this research project. S.J. and A.M. built the experimental setups, collected the measurements, and analyzed the results with support and supervision from M.P., R.I.W., T.J.D., R.M.S., and D.R., and particularly, derived the simulation results under the supervision of M.C., S.J.,

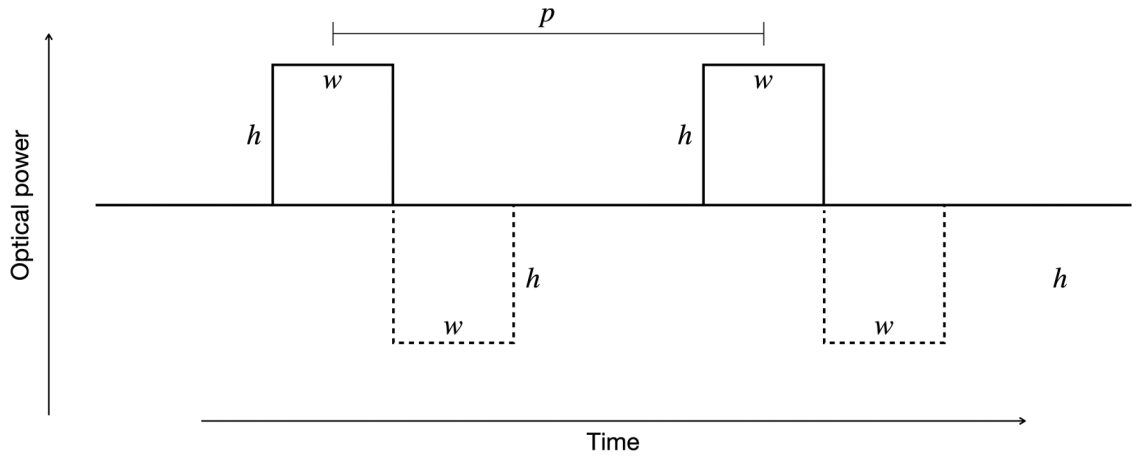


FIG. 7. Modulation example of Eve's attack. The schematics describe the nominal input provided to the Int Mod in Fig. 2. Solid lines refer to an "Up" configuration where Eve periodically increases the input power, whereas by appending to this modulation a symmetric reduction in power to keep the average power constant (dashed lines), Eve can prepare the "Up-to-Down" configuration. The actual modulated signal in the latter case is displayed in the first plot of Fig. 3(a). The "Down" and "Down-to-Up" configurations are obtained by considering vertical mirrored versions of these schemes. For each configuration we test different values of the pulse height h, the pulse width w and the modulation period p.

A.M., and R.I.W. wrote the manuscript, with all authors contributing to its improvement and the verification of the results.

The authors declare no competing interests.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain commercially sensitive information. The data are available from the authors upon reasonable request.

APPENDIX: METHODS

1. Experimental details of the fast intensity modulation attack

In this section we present the details of our investigation on the modulation parameters for Eve, and discuss how for every configuration a sufficiently high repetition rate deceives slowly integrating power meters, which therefore constitute insecure watchdogs.

Figure 7 reports an example of input pattern for Eve's intensity modulator for the case of a "Up" modulation (that is, fast, periodic increases in the optical power of the reference signal), as well as for the "Up-to-Down" modulation introduced in Fig. 3. The vertically mirrored patterns yield the "Down" and the "Down-to-Up" configurations, respectively. For each of the four configurations, we tested input and output statistics by changing the attack repetition rate (corresponding to $1/p$ from Fig. 7) and the modulation width w and height h. Crucially, as noted in Fig. 3, for both the "Up-to-Down" and "Down-to-Up" modulations of the reference signal, the locked light at Alice's output always results in a "Down-to-Up" configuration. This phenomenon also occurs for the "Up" and "Down" schemes, but notably in these latter cases the induced peak above the baseline is much smaller than the inflection below the baseline, resulting in a worse quality of the injection and overall reduction of the mean photon density. As this will affect performance rather than security, here we focus on the analysis for the "Up-to-Down" and "Down-to-Up" configurations, which allow us instead to induce a significant transient increase in the optical power.

For the repetition rate of the attack we tested values ranging from 100 MHz to 5 GHz, comparable with the operational rate of state-of-the-art QKD schemes. For all these options it holds $p \gg w$, meaning that the produced pulse shape is not altered, and no appreciable changes have been observed in the detection statistics. Therefore, in this analysis we consider a modulation rate of 1 GHz. As for the modulation width, we analyzed the produced pulse shapes while sweeping over the range 50 ps–250 ps. We observed that a short modulation is preferable as it allows for a steeper transient between low and high levels of the photon density, which ultimately results in wider and taller peaks above the baseline. Therefore, results in this section refer to the setting choice $w = 50$ ps, corresponding to the minimal resolution of our controls.

Finally, we tested various magnitudes of the modulation height h. Increasing this parameter directly impacts the modulation amplitude of the photon density. As this quantity is the most objective and meaningful metric for a device-agnostic analysis, figures in this paper show various experimental results as a function of Eve's normalized intensity modulation, corresponding to the peak-to-peak amplitude of the SNSPDs results displayed, for example, in Fig. 3(a).

The full experimental results for the slow-integrating power meters are displayed in the Supplemental Material [39], where figures report the normalized average power and normalized power fluctuations detected by the devices, respectively, and compare them with the attack-free scenario. The results discussed in the main text were obtained with an input power of 5 mW before the intensity modulator shown in Fig. 2.

Statistics are computed over $O(10^5)$ samples for each configuration. We observe that regardless of the attack intensity

and integration time, power meters fail to detect anomalies in the observed quantities, thus mistakenly assuming a safe scenario when Alice's laser is under a modulation attack. This motivates us to introduce surveillance procedures, namely the countermeasures introduced in Sec. III, so as to guarantee the implementation security of OIL-based TF-QKD against this novel class of potential threats.

2. Simulation details

Here we present the details of the SKR simulations displayed in Fig. 5. Our analysis is based on the protocol description and security proof for the "Sending-or-Not-Sending" (SNS) TF-QKD scheme introduced in Ref. [9] and the simulations provided in Ref. [11]. For simplicity, we consider the asymptotic case of infinitely many signals sent and infinitely many decoy intensities: the former implies that we can consider the key basis (Z basis) to be chosen with probability $P_Z \to 1$, while the second implies that all the required yields can be computed exactly. In addition, we consider a standard channel model, where losses are modeled as a beam splitter whose transmittance matches the one of the optical fiber.

Let η denote the total channel transmittance for a given distance between Alice and Bob and let $t = \sqrt{\eta}$ be the transmittance of the symmetric channels connecting each of the legitimate parties to Charlie. Here we consider a standard fiber loss coefficient $\alpha = 0.2$ dB/km and detectors with dark count probability $p_d = 10^{-8}$ and perfect detection efficiency (note that a finite detection efficiency could be incorporated in the definition of η). Following the definition of key events in Ref. [9], the vacuum and single-photon yields are found to be, respectively [9,11],

$$s_0 = 2p_d(1-p_d), \tag{A1}$$

$$s_1 = 2(1-p_d)\left[p_d(1-t) + \frac{t}{2}\right]. \tag{A2}$$

Let μ denote the intensity of the coherent state in the Z basis and ϵ the probability of Alice and Bob sending such coherent state (instead of vacuum) when they select a key round. For the total yield of correct and erroneous bits in the Z basis, we have

$$S_z^{\text{corr}} = 4\epsilon(1-\epsilon)(1-p_d)e^{-\mu t}(p_d + e^{\mu t/2} - 1), \tag{A3}$$

$$S_z^{\text{err}} = 2\epsilon^2(1-p_d)e^{-2\mu t}(p_d + e^{\mu t} I_0[\mu t \cos(\Delta\varphi)] - 1) + 2(1-\epsilon)^2 p_d(1-p_d), \tag{A4}$$

where $\Delta\varphi$ quantifies the polarization misalignment and $I_0(x)$ denotes the modified Bessel function of the first kind and order zero. It follows that for the total Z-basis yield and QBER we have:

$$S_z = S_z^{\text{corr}} + S_z^{\text{err}}, \quad E_z = \frac{S_z^{\text{err}}}{S_z}. \tag{A5}$$

We remark that in the SNS-TF-QKD protocol the notion of a valid X-basis event, required to estimate the phase error rate using decoy states, is different from that of a Z-basis event. In detail, considering a decoy-state intensity ν, we have that

$$S_x^{\text{corr}} = (1-p_d)[e^{-\nu t(1-\cos(\Delta\varphi)\cos(\Delta\theta))} - (1-p_d)e^{-2\nu t}], \tag{A6}$$

$$S_x^{\text{err}} = (1-p_d)[e^{-\nu t(1+\cos(\Delta\varphi)\cos(\Delta\theta))} - (1-p_d)e^{-2\nu t}], \tag{A7}$$

being $\Delta\theta$ the phase misalignment. The total X-basis yield and QBER follow as

$$S_x = S_x^{\text{corr}} + S_x^{\text{err}}, \quad E_x = \frac{S_x^{\text{err}}}{S_x}. \tag{A8}$$

From these quantities, an upper bound on the phase error rate can be found to be [9]:

$$e_{\text{ph}} \leqslant \bar{e}_{\text{ph}} = \frac{S_x E_x - \frac{1}{2}s_0 e^{-2\nu}}{2\nu e^{-2\nu} s_1}. \tag{A9}$$

The above bound holds for any value of ν and one can numerically check that it gets tighter as $\nu \ll 1$. Therefore, in our calculations we considered the fixed value of $\nu = 10^{-6}$.

Finally, the secret key rate R can be obtain as

$$R = 2\epsilon(1-\epsilon)\mu e^{-\mu} s_1[1 - h(\bar{e}_{\text{ph}})] - f_E\, S_z\, h(E_z), \tag{A10}$$

where $h(x)$ denotes the binary entropy of x and $f_E = 1.16$ is the error correction inefficiency.

As mentioned in the caption of Fig. 5, before actually running the protocol Alice and Bob establish the optimal key parameters

$$(\tilde{\epsilon}, \tilde{\mu}) := \underset{\epsilon,\mu}{\text{argmax}}\, R \tag{A11}$$

for a given set of system parameters η, p_d, $\Delta\varphi$, and $\Delta\theta$. All curves in Fig. 5 refer to the optimal case $\Delta\varphi = \Delta\theta = 0$, but no significant difference has been observed when comparing to higher misalignment cases.

Due to the attack, although Alice and Bob expect to prepare Z states with intensity $\tilde{\mu}$ they actually prepare states with intensity $\tilde{\mu}' = \kappa\mu$, for some $\kappa > 0$ ($\kappa = 1.51$ in our case, from Fig. 3). With a noiseless channel model, the Z-basis statistics that Alice and Bob actually observe are obtained through Eqs. (A3), (A4), and (A5) with $\mu = \tilde{\mu}'$. This shows that within this framework the attack has a direct impact on the Z-basis QBER, while its effect on the phase error rate is negligible as $\nu \approx 0$.

The actual secret key rate guaranteed by the security proof in this case is obtained by plugging in the actual intensity sent $\tilde{\mu}'$ and the actual bit error rate E_z in Eq. (A10), together with the sending probability $\tilde{\epsilon}$, which is unchanged.

Nevertheless, if Alice and Bob are unaware of the attack, they believe they are sending coherent states with the average intensity $\tilde{\mu}$ and observe the actual bit error rate E_z, as determined from Charlie's announcements. As a result, they compute their expected key rate by substituting $\tilde{\mu}$ and E_z into Eq. (A10), which leads to an erroneous overestimation.

[1] C. H. Bennett and G. Brassard, Quantum cryptography: Public key distribution and coin tossing, Theor. Comput. Sci. **560**, 7 (2014).

[2] A. K. Ekert, Quantum cryptography based on Bell's theorem, Phys. Rev. Lett. **67**, 661 (1991).

[3] S. Pirandola, U. L. Andersen, L. Banchi, M. Berta, D. Bunandar, R. Colbeck, D. Englund, T. Gehring, C. Lupo, C. Ottaviani, J. L. Pereira, M. Razavi, J. Shamsul Shaari, M. Tomamichel, V. C. Usenko, G. Vallone, P. Villoresi, and P. Wallden, Advances in quantum cryptography, Adv. Opt. Photon. **12**, 1012 (2020).

[4] H.-K. Lo, M. Curty, and K. Tamaki, Secure quantum key distribution, Nat. Photon. **8**, 595 (2014).

[5] M. Sasaki, M. Fujiwara, H. Ishizuka, W. Klaus, K. Wakui, M. Takeoka, S. Miki, T. Yamashita, Z. Wang, A. Tanaka, *et al.*, Field test of quantum key distribution in the tokyo QKD network, Opt. Express **19**, 10387 (2011).

[6] V. Martin, J. P. Brito, L. Ortíz, R. Méndez, J. Buruaga, R. Vicente, A. Sebastian-Lombrana, D. Rincon, F. Perez, C. Sanchez, *et al.*, MadQCI: a heterogeneous and scalable SDN-QKD network deployed in production facilities, npj Quantum Inf. **10**, 80 (2024).

[7] H.-Z. Chen, M.-H. Li, Y. Z. Wang, Z.-G. Zhao, C. Ye, F. L. Li, Z. Chen, S.-L. Han, B. Tang, Y. J. Miao, *et al.*, Implementation of carrier-grade quantum communication networks over 10000 km, npj Quantum Inf. **11**, 137 (2025).

[8] M. Lucamarini, Z. L. Yuan, J. F. Dynes, and A. J. Shields, Overcoming the rate-distance limit of quantum key distribution without quantum repeaters, Nat. (London) **557**, 400 (2018).

[9] X.-B. Wang, Z.-W. Yu, and X.-L. Hu, Twin-field quantum key distribution with large misalignment error, Phys. Rev. A **98**, 062323 (2018).

[10] H. Xu, Z.-W. Yu, C. Jiang, X.-L. Hu, and X.-B. Wang, Sending-or-not-sending twin-field quantum key distribution: Breaking the direct transmission key rate, Phys. Rev. A **101**, 042330 (2020).

[11] M. Curty, K. Azuma, and H.-K. Lo, Simple security proof of twin-field type quantum key distribution protocol, npj Quantum Inf. **5**, 64 (2019).

[12] M. Curty, T. Moroder, X. Ma, H.-K. Lo, and N. Lütkenhaus, Upper bounds for the secure key rate of the decoy-state quantum key distribution, Phys. Rev. A **79**, 032335 (2009).

[13] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, Fundamental limits of repeaterless quantum communications, Nat. Commun. **8**, 15043 (2017).

[14] M. Minder, M. Pittaluga, G. L. Roberts, M. Lucamarini, J. F. Dynes, Z. L. Yuan, and A. J. Shields, Experimental quantum key distribution beyond the repeaterless secret key capacity, Nat. Photon. **13**, 334 (2019).

[15] M. Pittaluga, M. Minder, M. Lucamarini, M. Sanzaro, R. I. Woodward, M.-J. Li, Z. Yuan, and A. J. Shields, 600-km repeater-like quantum communications with dual-band stabilization, Nat. Photon. **15**, 530 (2021).

[16] J.-P. Chen, C. Zhang, Y. Liu, C. Jiang, D.-F. Zhao, W.-J. Zhang, F.-X. Chen, H. Li, L.-X. You, Z. Wang, Y. Chen, X.-B. Wang, Q. Zhang, and J.-W. Pan, Quantum key distribution over 658 km fiber with distributed vibration sensing, Phys. Rev. Lett. **128**, 180502 (2022).

[17] S. Wang, Z.-Q. Yin, D.-Y. He, W. Chen, R.-Q. Wang, P. Ye, Y. Zhou, G.-J. Fan-Yuan, F.-X. Wang, Y.-G. Zhu, P. V. Morozov, A. V. Divochiy, Z. Zhou, G.-C. Guo, and Z.-F. Han, Twin-field quantum key distribution over 830-km fibre, Nat. Photon. **16**, 154 (2022).

[18] Y. Liu, W.-J. Zhang, C. Jiang, J.-P. Chen, C. Zhang, W.-X. Pan, D. Ma, H. Dong, J.-M. Xiong, C.-J. Zhang, H. Li, R.-C. Wang, J. Wu, T.-Y. Chen, L. You, X.-B. Wang, Q. Zhang, and J.-W. Pan, Experimental twin-field quantum key distribution over 1000 km fiber distance, Phys. Rev. Lett. **130**, 210801 (2023).

[19] M. Pittaluga, Y. S. Lo, A. Brzosko, R. I. Woodward, D. Scalcon, M. S. Winnel, T. Roger, J. F. Dynes, K. A. Owen, S. Juárez, *et al.*, Long-distance coherent quantum communications in deployed telecom networks, Nat. (London) **640**, 911 (2025).

[20] J.-P. Chen, C. Zhang, Y. Liu, C. Jiang, W.-J. Zhang, Z.-Y. Han, S.-Z. Ma, X.-L. Hu, Y.-H. Li, H. Liu, F. Zhou, H.-F. Jiang, T.-Y. Chen, H. Li, L.-X. You, Z. Wang, X.-B. Wang, Q. Zhang, and J.-W. Pan, Twin-field quantum key distribution over a 511 km optical fibre linking two distant metropolitan areas, Nat. Photon. **15**, 570 (2021).

[21] H. Liu, C. Jiang, H.-T. Zhu, M. Zou, Z.-W. Yu, X.-L. Hu, H. Xu, S. Ma, Z. Han, J.-P. Chen, Y. Dai, S.-B. Tang, W. Zhang, H. Li, L. You, Z. Wang, Y. Hua, H. Hu, H. Zhang, F. Zhou, *et al.*, Field test of twin-field quantum key distribution through sending-or-not-sending over 428 km, Phys. Rev. Lett. **126**, 250502 (2021).

[22] C. Clivati, A. Meda, S. Donadello, S. Virzì, M. Genovese, F. Levi, A. Mura, M. Pittaluga, Z. Yuan, A. J. Shields, M. Lucamarini, I. P. Degiovanni, and D. Calonico, Coherent phase transfer for real-world twin-field quantum key distribution, Nat. Commun. **13**, 157 (2022).

[23] L. Zhou, J. Lin, Y. Jing, and Z. Yuan, Twin-field quantum key distribution without optical frequency dissemination, Nat. Commun. **14**, 928 (2023).

[24] F. Xu, X. Ma, Q. Zhang, H.-K. Lo, and J.-W. Pan, Secure quantum key distribution with realistic devices, Rev. Mod. Phys. **92**, 025002 (2020).

[25] V. Zapatero, Á. Navarrete, and M. Curty, Implementation security in quantum key distribution, Adv. Quantum Technol. **8**, 2300380 (2025).

[26] N. Jain, B. Stiller, I. Khan, D. Elser, C. Marquardt, and G. Leuchs, Attacks on practical quantum key distribution systems (and how to prevent them), Contemp. Phys. **57**, 366 (2016).

[27] V. Makarov, A. Abrikosov, P. Chaiwongkhot, A. K. Fedorov, A. Huang, E. Kiktenko, M. Petrov, A. Ponosova, D. Ruzhitskaya, A. Tayduganov, D. Trefilov, and K. Zaitsev, Preparing a commercial quantum key distribution system for certification against implementation loopholes, Phys. Rev. Appl. **22**, 044076 (2024).

[28] Federal Office for Information Security (BSI), Implementation Attacks against QKD Systems, Tech. Rep., Federal Office for Information Security (BSI) (2023), https://www.bsi.bund.de/EN/Service-Navi/Publikationen/Studien/QKD-Systems/Implementation_Attacks_QKD_Systems_node.html.

[29] H.-K. Lo, M. Curty, and B. Qi, Measurement-device-independent quantum key distribution, Phys. Rev. Lett. **108**, 130503 (2012).

[30] J. Ye, J.-L. Peng, R. J. Jones, K. W. Holman, J. L. Hall, D. J. Jones, S. A. Diddams, J. Kitching, S. Bize, J. C. Bergquist, L. W. Hollberg, L. Robertsson, and L.-S. Ma, Delivery of high-stability optical and microwave frequency standards over an optical fiber network, J. Opt. Soc. Am. B **20**, 1459 (2003).

[31] L. C. Comandar, M. Lucamarini, B. Fröhlich, J. F. Dynes, Z. Yuan, and A. J. Shields, Near perfect mode overlap between independently seeded, gain-switched lasers, Opt. Express **24**, 17849 (2016).

[32] T. K. Paraïso, R. I. Woodward, D. G. Marangon, V. Lovic, Z. Yuan, and A. J. Shields, Advanced laser technology for quantum communications (tutorial review), Adv. Quantum Technol. **4**, 2100062 (2021).

[33] H. Du, T. K. Paraiso, M. Pittaluga, Y. S. Lo, J. A. Dolphin, and A. J. Shields, Twin-field quantum key distribution with optical injection locking and phase encoding on-chip, Optica **11**, 1385 (2024).

[34] Q. Peng, J.-P. Chen, T. Xing, D. Wang, Y. Wang, Y. Liu, and A. Huang, Practical security of twin-field quantum key distribution with optical phase-locked loop under wavelength-switching attack, npj Quantum Inf. **11**, 7 (2025).

[35] M. Lucamarini, I. Choi, M. B. Ward, J. F. Dynes, Z. L. Yuan, and A. J. Shields, Practical security bounds against the Trojan-horse attack in quantum key distribution, Phys. Rev. X **5**, 031030 (2015).

[36] H.-K. Lo, X. Ma, and K. Chen, Decoy state quantum key distribution, Phys. Rev. Lett. **94**, 230504 (2005).

[37] S. Sajeed, I. Radchenko, S. Kaiser, J.-P. Bourgoin, A. Pappa, L. Monat, M. Legré, and V. Makarov, Attacks exploiting deviation of mean photon number in quantum key distribution and coin tossing, Phys. Rev. A **91**, 032326 (2015).

[38] A. Huang, A. Navarrete, S.-H. Sun, P. Chaiwongkhot, M. Curty, and V. Makarov, Laser-seeding attack in quantum key distribution, Phys. Rev. Appl. **12**, 064043 (2019).

[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/71m5-3c5n for the slow-integrating power meters measurements against the fast intensity modulation (FIM) attack.

[40] H. Tan, M. Petrov, W. Zhang, L. Han, S.-K. Liao, V. Makarov, F. Xu, and J.-W. Pan, Wide-spectrum security of quantum key distribution, PRX Quantum **6**, 040331 (2025).

[41] A. Rogalski, *Infrared Detectors*, 2nd ed. (CRC Press, Boca Raton, 2011).

[42] C. W. Carr, H. B. Radousky, and S. G. Demos, Wavelength dependence of laser-induced damage: Determining the damage initiation mechanisms, Phys. Rev. Lett. **91**, 127402 (2003).

[43] J. C. Palais, *Fiber Optic Communications*, 5th ed. (Pearson/Prentice Hall, Hoboken, 2005).

[44] G. Currás-Lorenzo, A. Navarrete, J. Núñez-Bon, M. Pereira, and M. Curty, Numerical security analysis for quantum key distribution with partial state characterization, Quantum Sci. Technol. **10**, 035031 (2025).

[45] K. Tamaki, M. Curty, and M. Lucamarini, Decoy-state quantum key distribution with a leaky source, New J. Phys. **18**, 065008 (2016).

[46] W. Wang, K. Tamaki, and M. Curty, Finite-key security analysis for quantum key distribution with leaky sources, New J. Phys. **20**, 083027 (2018).

[47] A. Navarrete and M. Curty, Improved finite-key security analysis of quantum key distribution against trojan-horse attacks, Quantum Sci. Technol. **7**, 035021 (2022).

[48] G. Currás-Lorenzo, M. Pereira, G. Kato, M. Curty, and K. Tamaki, Security framework for quantum key distribution with imperfect sources, Optica Quantum **3**, 525 (2025).

[49] X. Sixto, Á. Navarrete, M. Pereira, G. Currás-Lorenzo, K. Tamaki, and M. Curty, Quantum key distribution with imperfectly isolated devices, Quantum Sci. Technol. **10**, 035034 (2025).