Securing Practical Quantum Communication Systems with Optical Power Limiters

Gong Zhang,[1,*] Ignatius William Primaatmaja,[2] Jing Yan Haw,[1] Xiao Gong,[1] Chao Wang,[1,†] and Charles Ci Wen Lim[1,2,‡]

[1] *Department of Electrical & Computer Engineering, National University of Singapore, Singapore*
[2] *Centre for Quantum Technologies, National University of Singapore, Singapore*

(Received 2 March 2021; accepted 1 June 2021; published 7 July 2021)

Controlling the energy of unauthorized light signals in a quantum cryptosystem is an essential criterion for implementation security. Here, we propose a passive optical power limiter device based on thermo-optical defocusing effects, providing a reliable power limiting threshold that can be readily adjusted to suit various quantum applications. In addition, the device is robust against a wide variety of signal variations (e.g., wavelength, pulse width), which is important for implementation security. We experimentally show that the proposed device does not compromise quantum communication signals. It only has a minimal impact (if not, negligible impact) on the intensity, phase, or polarization degrees of freedom of the photons, thus making it suitable for general communication purposes. To show its practical utility for quantum cryptography, we demonstrate and discuss three potential applications: (1) measurement-device-independent quantum key distribution with enhanced security against a general class of Trojan-horse attacks, (2) using the power limiter as a countermeasure against bright illumination attacks, and (3) the application of power limiters to potentially enhance the implementation security of plug-and-play quantum key distribution.

DOI: 10.1103/PRXQuantum.2.030304

I. INTRODUCTION

Quantum key distribution (QKD) enables two remote network users to exchange provably secure keys when it is implemented faithfully [1–3]. To ensure implementation security, the research community has been focusing on the security of practical systems in recent years, developing methods to narrow the gap between the theory and practice of QKD. On the theoretical side, robust QKD protocols have been proposed, which not only make practical systems more secure against device imperfections but also easier to calibrate and validate in practice (since fewer assumptions are required). On the experimental side, efforts have been focused on tackling quantum side channels and a wide variety of countermeasures have been proposed and developed [3,4].

Trojan-horse attacks (THAs) [5,6] represent one of the biggest threats to QKD security. These attacks aim to steal the secret key information via the injection of unauthorized light pulses, seeking to carry critical modulation information out of the transmitters. More specifically, in these attacks, the adversary (henceforth called Eve) injects bright light pulses into the transmitter and collects the reflected light pulses. Consequently, this allows Eve to learn some information about the secret key. It has been shown that these kinds of attacks can be readily implemented using standard optical methods [5–8]. To mitigate this issue, one can use specialized security analyses to include security against specific types of THAs, for instance, by modeling the unauthorized input light pulses as coherent states. Then, under the assumption that the energy of the reflected light pulses is bounded, one can compute the secret key rate [9,10].

The bright illumination attacks are another particularly powerful class of side-channel attacks. These include laser damage attacks [11–13] and blinding attacks [14–16]. In these attacks, bright light pulses are used to control QKD devices by exploiting their implementation knowledge. Consequently, these allow Eve to avoid eavesdropping detection and hence security is no longer guaranteed. Fortunately, there exist countermeasures that are fairly effective against such attacks [16,17] and innovative QKD protocols that are completely immune against detection-side-channel attacks are known as well, e.g.,

*zhanggong@nus.edu.sg
†wang.chao@nus.edu.sg
‡charles.lim@nus.edu.sg

see measurement-device-independent QKD (MDI QKD) [18,19].

Based on the above, it can therefore be said that the injection of (unauthorized) bright light pulses into quantum communication systems is a catalyst for side-channel attacks. This is not so surprising since the presence of bright light pulses essentially breaks one of the most important assumptions of quantum cryptography—that the energy of the underlying quantum signals is at the single-photon level (or sufficiently small). To overcome these potential loopholes, one promising solution is to limit the energy of incoming light. Indeed, if this is achieved, one can be sure that the QKD system is operating at the single-photon level and the energy of any outgoing light pulse is bounded as well. Consequently, this will allow the system to operate faithfully in the quantum regime.

In practice, this solution would mean introducing a kind of *quantum power limiting* device into the QKD system. Based on current research on side-channel attacks, we believe that an ideal quantum power limiter should possess the following properties: (a) able to provide a reliable and adjustable photon energy limit down to the order of a few photons to hundreds of photons for each quantum state, (b) have a minimum insertion loss if the input power is below the threshold and stop the transmission or maintain at the threshold power once the input power exceeds the threshold, and (c) the power limiting effects are independent of other physical degrees of freedoms, such as frequency and polarization. In terms of practical considerations, the power limiter device should also be cost effective, passive, and easily replaceable (if it cannot recover to its normal state after exposure to strong light).

Here, we propose and demonstrate a practical quantum power limiter that can secure a broad class of QKD setups [20]. The device is based on the thermo-optical defocusing effect, which effectively bounds the output optical power by some predetermined threshold. By modeling the system using a set of physically relevant assumptions, we show that the output-input optical power relation of the proposed device can be precisely controlled by changing the system parameters, e.g., the length of the prism and the diaphragm width. Consequently, this allows us to tailor the device to different quantum cryptographic applications. The feasibility and performance of our proposed power limiter device are confirmed using COMSOL (a multiphysics simulation software) and experimental data.

The paper is organized as follows. In Sec. II, we first present the design details and the modeling of our power limiter. Thereafter, simulation and experimental results are illustrated. In Sec. III, we discuss the potential implementation loopholes and the robustness of the proposed power limiter. In Sec. IV, we experimentally verify that the power limiter is essentially transparent to standard quantum encoding choices such as intensity, phase, and polarization degrees of freedom. In Sec. V, we illustrate the broad utility of the proposed power limiter over three different QKD systems. In the first application, we provide a general security analysis of MDI QKD that allows Eve to inject in any kind of state in a given Trojan-horse optical mode. After that, a detailed study on the application of our power limiter in MDI QKD is presented, followed by the simulation results. In the second and third applications, we discuss how the proposed power limiter could be utilized to deter bright illumination attacks and to enhance the implementation security of plug-and-play QKD [21–23]. We conclude in Sec. VI.

II. OPTICAL POWER LIMITER DESIGN

Our power limiter design is shown in Fig. 1. The input light and output light are collimated using a pair of fiber collimators. An acrylic prism is placed along the optical path as the core part of our proposal, whose negative thermo-optical coefficient (TOC) dn/dT is exploited, where n is the refractive index and T is the temperature. Note that any material with negative TOC could be used with similar analysis. The absorption of input light generates a heat gradient inside the prism, which is then converted to a refractive index gradient accordingly. The negative TOC leads to a relatively smaller reflective index at the center of the prism, resulting in the whole optical architecture working as a concave lens and diverging the transmitting light, as shown in the inset of Fig. 1. By adding a diaphragm (also known as a pinhole or an iris) with a customizable aperture width (referred to as the

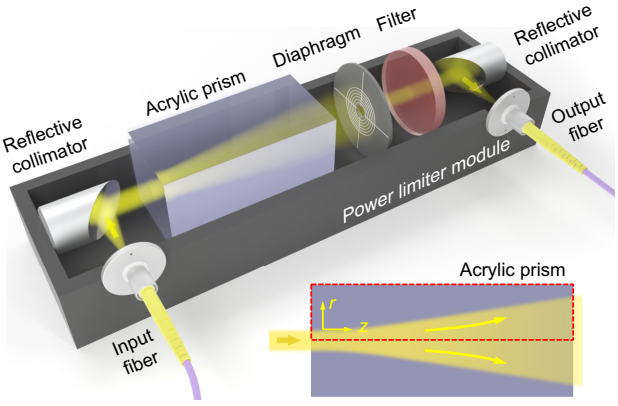

FIG. 1. Schematic of the power limiter design. An acrylic prism is used as the active medium. When the absorbed energy introduces temperature gradients inside the prism, the input collimated Gaussian beam diverges due to the thermo-optical defocusing effect. A diaphragm is placed after the prism to control the collectible optical power. The optical filter restricts the working wavelength range for security analysis. The inset is the top view of the acrylic prism and the diverged Gaussian beam. Owing to the isotropic nature of acrylic, both the optical and thermal responses are assumed to be axially symmetric along the optical axis.

"diaphragm width" below), the amount of output power can be suitably controlled. An optical filter is then introduced to restrict the working wavelength range of the device for security analysis, which we discuss in detail in Sec. III. We remark that all components used here are cost effective and commercially available.

The mechanism of thermal optical defocusing and related power limiting phenomena have been widely studied in both theory and experiments [24–26]. In our case, we first simulate the temperature and electric field distribution inside a 100 mm acrylic prism with 7.9 mW input power using COMSOL, whose results are shown in Figs. 2(a) and 2(b). The simulation results indicate a distinct temperature distribution inside the medium, and a clear divergence of the light field. Towards a better quantitative understanding, we model our power limiter design by balancing the optical absorption and the heat transfer inside the prism under the steady-state condition [27]

$$\alpha I = -\frac{k}{r}\frac{\partial}{\partial r}\left(r\frac{\partial T}{\partial r}\right), \tag{1}$$

where α is the absorption coefficient of the material, I represents the input light power density, T is the temperature, and k is the thermal conductivity. Here we assume that the light propagates along the z direction, the temperature gradient in the z direction is negligible, and that the radiative and convective heat transfer is minimal. If we further assume that the light follows a Gaussian profile, The steady-state laser radiation intensity at position (r,z) can be solved as [27]

$$I(r,z) = I(r,0) \exp\left[-\alpha z + \frac{(\partial n/\partial T)P_0 e^{-r^2/a^2}[z-(1-e^{-\alpha z})/\alpha]}{\pi k n a^2}\right], \tag{2}$$

where the input intensity $I(r,0) = P_0 e^{-r^2/a^2}/\pi a^2$, a is the radius where the light intensity drops to $1/e$ of its axial value, and P_0 is the incident laser power. The output optical power can be obtained by integrating the light intensity over a certain area that depends on the position (prism length) and the diaphragm width.

The maximum output power (defined as the power limiting threshold) and the insertion loss at different prism lengths and diaphragm widths are shown in Figs. 2(c) and 2(d). Since a larger prism length leads to a greater photon absorption as well as a larger light divergence, a higher insertion loss and a smaller power limiting threshold can thus be expected. Likewise, a smaller diaphragm collects less photon energy, which also results in a higher insertion loss and a smaller power limiting threshold. Therefore, depending on the application, it is possible to choose a set of parameters that balance the insertion loss and power limiting threshold that meet system requirements.

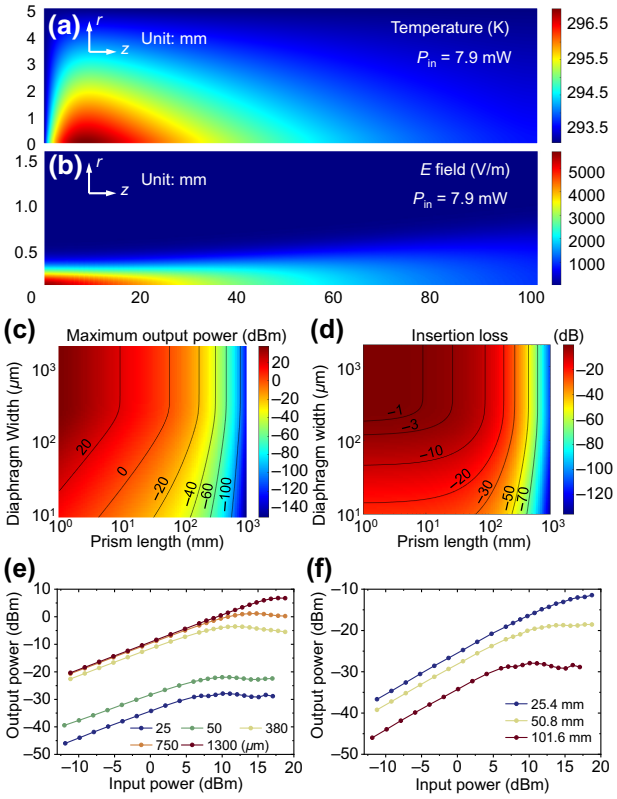

FIG. 2. Simulated (a) temperature and (b) E-field distribution in the region marked with the dashed box in Fig. 1 using a two-dimensional model in COMSOL. The results clearly indicate the high-temperature gradient in the r direction and the gradual divergence of the Gaussian profile in the r direction. Simulated (c) maximum output power and (d) insertion loss at different diaphragm widths and prism lengths using Eq. (2), where $\alpha = 25.95\ \mathrm{m}^{-1}$ (measured value), $\mathrm{TOC} = -1.3\times10^{-4}\ \mathrm{K}^{-1}$ [58], $n = 1.47$ [29], $k = 0.19\ \mathrm{Wm}^{-1}\mathrm{K}^{-1}$ [30], $a = 140\ \mu$m. (e) Experimental output-input power relationship at different diaphragm widths with the same prism length of 101.6 mm. The results indicate that the smaller the diaphragm width, the lower the output power threshold. Also, with a diaphragm width larger than the beam width, the insertion loss remains minimum. (f) Experimental output-input power relationship at different prism lengths but with the same diaphragm width of 25 μm. Longer prism lengths could provide a lower output power threshold, but the insertion loss could be higher. Both the simulation and experimental results confirm the power limiting effect of our design, and an adjustable power limiting threshold is feasible.

Note here that the Gaussian profile assumption only holds when the beam divergence is relatively small. Thus, the analytical model may only be able to provide a quick guide for parameter selection. Hence, experiments are conducted to verify the feasibility of our proposal.

A proof-of-concept experiment is performed using a simplified version of Fig. 1. A collimator is used for light coupling from a single-mode optical fiber to free space. Here a transmissive collimator based on a gradient-index

lens is used in the setup for feasibility demonstration, and it can be conveniently replaced by reflective collimators to ensure the proper functioning over a wide range of wavelengths for security reasons (see Sec. III for details). Then the Gaussian beam with a beamwidth of 0.4 mm is directed into the acrylic prism. Three acrylic prisms with lengths 25.4, 50.8, and 101.6 mm are tested. The output light can then be collected after the diaphragm. Diaphragm widths of 25, 50, 380, 750, and 1300 μm are used in our experiments. Figure 2(e) shows the measured output-input relationship at different diaphragm widths and the same prism length of 101.6 mm, while Fig. 2(f) shows the result at different prism lengths with the same diaphragm width of 25 μm. The results clearly show the power limiting effect in various conditions. The output power linearly increases with the input power at the low power region. As the input power further increases, the output power increases slowly and is finally limited to a certain threshold.

Besides, the experimental results verify that the power limiting feature of our proposal can be readily adjusted by modifying the prism length and diaphragm width. Among all of our system configurations, the lowest power limiting threshold of -27.9 dBm is measured, with an insertion loss of -34.0 dB, when a 101.6 mm prism and 25 μm diaphragm are chosen. Similarly, a lower insertion loss of -5.1 dB can be obtained, together with a 10.3 dBm output power limiting threshold, when a 50.8 mm prism and 750 μm diaphragm are used. For different applications, one can expect different requirements for power limit devices. For example, for protecting transmitters against THAs, the insertion loss of the power limiter is less concerning since we can always adjust the optical attenuators to generate expected quantum states. While in order to protect receivers from bright illumination attacks, the insertion loss of the device can be a critical factor in system performance. Thus, we would imagine a customized power limiter configuration is required for different application scenarios.

III. ROBUSTNESS AGAINST POTENTIAL IMPLEMENTATION LOOPHOLES

The above analyses so far only show the feasibility of the proposed power limiter under a steady-state condition. Below, we analyze the robustness of the proposed device against potential implementation loopholes that could happen via the variation of standard optical properties.

One important consideration is the finite response time of the proposed device. To investigate this property, we install an electronic variable optical attenuator after a continuous-wave (CW) laser source to create a laser pulse with a relatively long pulse width and measure the output response of the power limiter. The experimental results are shown in Fig. 3(a), where a 101.6 mm prism is used with a 750 μm diaphragm. These parameters are selected to provide a strong power limiting effect while maintaining a relatively low insertion loss. The settling time of our power limiter is measured to be 300 ms. We observe that the peak output power close to the starting time can be a few times higher than the steady-state output power (which happens after about 300 ms). Crucially, this suggests that one could exploit the finite response time of the power limiter to breach the desired energy threshold.

As we show in Sec. V A, however, the information leakage due to THAs can, in fact, be bounded using only the average energy constraint (integrated over the finite response time); it is not necessary to bound the maximum (peak) energy for security. Thus, in the experiment, we study the average output optical power at a constant-energy pulse input but with different duty cycles. The time-domain results are shown in Fig. 3(b), where a 101.6 mm prism is used with a 25 μm diaphragm. Here, based on our analysis in Sec. II, the longest prism and the smallest diaphragm width are chosen to obtain the strongest power limiting effect, which could offer a better comparison between CW light input and pulsed light inputs. The input

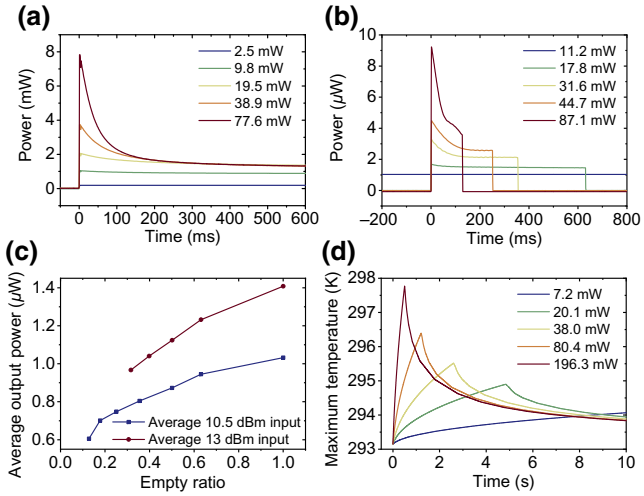

FIG. 3. (a) Experimental output response of the power limiter at different input optical powers. A 101.6 mm prism is used with a 750 μm diaphragm. The power limiting effect has a long settling time of around 300 ms. (b) Experimental output response of the power limiter with constant-energy pulse input. A 101.6 mm prism is used with a 25 μm diaphragm. The peak power and duty cycle are selected to maintain the same energy per pulse or the same average input power of 10.5 dBm. (c) The average output power at different duty cycles with average input powers of 10.5 and 13 dBm shows that the maximum output power occurs at a duty cycle of 1, i.e., CW input. (d) The COMSOL simulated maximum temperature inside the prism with constant pulse energy of 97.8 mJ and pulse widths of about 0.5 to 13.5 s, which correspond to an average input power of 9.78 mW in a 10 s time span. The result shows that higher peak power heats the prism faster and reaches a higher temperature. Thus, a higher thermo-optical effect can be expected.

laser pulse is modulated at 1 Hz frequency with average input powers of 10.5 and 13 dBm. The corresponding average output power is shown in Fig. 3(c). The results indicate that the average output power is higher at a larger duty cycle. The maximum appears at a duty cycle of 1, i.e., CW light input. In other words, given a fixed average input power, the CW input gives the largest averaged output power, where Eve is getting the most amount of information about the transmitter. As such, we will be using the power limiting threshold obtained under the CW Trojan-horse input assumption for the THA analysis; see Sec. V A. To explain this effect, we study the temperature response inside the medium under a constant-energy pulse input with different peak powers and different duty cycles using COMSOL. The results are shown in Fig. 3(d). The simulation results indicate that a higher input peak power will lead to a higher maximum temperature, even with the same amount of average power. Therefore, a higher refractive index gradient and larger divergence of the input laser are expected with a higher instantaneous power of the input light, leading to a larger thermo-optical defocusing effect and consequently a lower output power.

Another possible attack is to try to change the power limiting threshold by varying the wavelength of the incoming light. This could allow Eve to send in brighter light pulses with a different wavelength. To investigate the possibility of such an attack, we analyze how different input wavelengths could affect the TOC and heat generation of the power limiter device.

Generally, the TOC can be modeled by [31,32]

$$\mathrm{TOC} = \frac{dn}{dT} = f\,[n(\lambda)](\Phi - \beta), \tag{3}$$

where $f\,[n(\lambda)]$ is defined as $(n^2-1)(n^2+2)/(6n)$, n is the reflective index, λ is the wavelength of input light, Φ is the electronic polarizability, and β is the volumetric expansion coefficient. In most polymers, the volumetric expansion coefficient is more dominant, i.e., $\Phi \ll \beta$, and hence the overall TOC is typically negative [31]. More importantly, note that the volumetric expansion coefficient is physically independent of the wavelength. As such, the wavelength dependency of TOC is only related to $f\,[n(\lambda)]$. The $f\,[n(\lambda)]$ for acrylic as a function of wavelength is shown in Fig. 4(a). The corresponding TOC change will introduce a small difference in the output power threshold calculation, as referenced to the power at 1550 nm, which is shown as the red curve in Fig. 4(a).

As for heat generation, it is related to the absorption loss of the material. A lower loss indicates that less optical energy is converted to heat energy, thereby resulting in a lower temperature gradient and a higher power limiting threshold. Based on this, a spectral filter with a large power handling capability can be applied to limit the transmission spectrum of the device, in which case the peak power (over the transmitted spectrum) is considered for the security analysis.

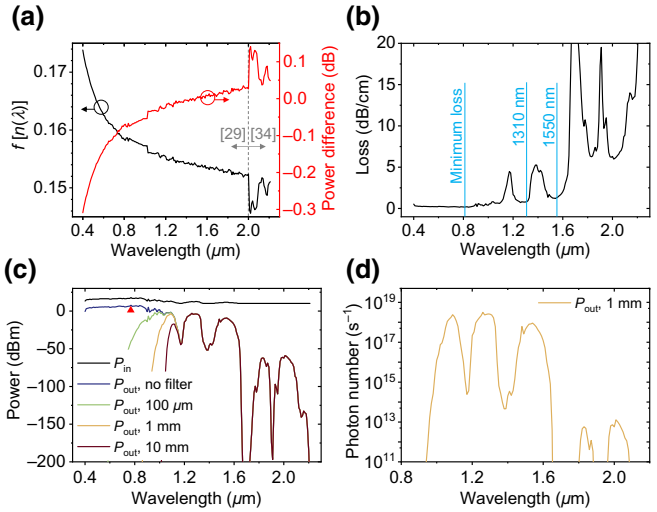

FIG. 4. (a) Calculated $f\,[n(\lambda)]$ from Eq. (3) and the output power threshold difference caused by the corresponding TOC change, as referenced to the value at 1550 nm. (b) Absorption loss spectrum of acrylic, taken from Refs. [29,34]. The losses at the 1310 and 1550 nm communication bands are marked. The minimum loss is about 0.15 dB/cm at visible wavelength. (c) The maximum output power of the power limiter and the corresponding input power with silicon absorber (visible light absorbing filter) of different lengths. The silicon absorber significantly reduces the power threshold at the visible wavelength. (d) The total maximum output photon number per second calculated from (c) a the silicon absorber thickness of 1 mm.

Considering optical fiber-based applications at 1550 nm, the optical fiber itself is, in fact, a bandpass filter for about 300–2100 nm wavelength, beyond which the transmission loss is higher than 100 dB/km [33]. Thus, by applying a secure fiber with adequate length (inside Alice's or Bob's apparatus), the light beyond this wavelength range can be suppressed to a negligible level. In this way, it is effective only to consider the wavelength dependency feature within this band.

For the material of our power limiter, acrylic, its absorption loss spectrum is shown in Fig. 4(b) with some low loss bands marked [29,34]. The loss is about 1.29 dB/cm at 1550 nm and 0.82 dB/cm at 1310 nm, which are standard communication bands. The loss below 1100 nm is even lower. The minimum occurs at about 800 nm with 0.15 dB/cm loss. Based on the absorption spectrum and considering a prism length of 10 cm with 750 μm diaphragm, the maximum output power spectrum and the corresponding input power are calculated based on Eq. (2), as shown in Fig. 4(c). The power threshold below 1100 nm is about 11 dB higher than the 1310 nm band and more than 17 dB higher than the 1550 nm band. Although a pessimistic power bound of about 8 dBm can be set and used as the

system power bound (marked as a red triangle), it is better to use an optical filter to block the light below 1100 nm wavelength. The silicon absorber can be a good candidate, which provides a stable and robust filtering performance. By adding a thin layer of silicon sheet after the power limiter, the output power can be significantly suppressed. As shown in Fig. 4(c), with only about 1 mm thick silicon, the maximum output power shifts back to the communication bands. The maximum output photon number per second can be further calculated, as shown in Fig. 4(d). The maximum output photon number per second appears at 1260 nm wavelength with a photon energy of 1.58×10^{-19} J; as such, this wavelength is considered in the security analysis using the worst-case approach.

Similarly, for other degrees of freedom, e.g., the state of polarization, the polymer acrylic used in our design inherently possesses isotropic behavior. Thus, by nature, it will not introduce any birefringence-related changes and is independent of the thermo-optical effect, which avoids introducing related loopholes to the system.

Another consideration is laser damage attacks [11–13]. Preliminary simulations indicate that the acrylic prism could be damaged with only about 400 mW of input power [35,36]. If the acrylic is damaged or burnt, the thermal defocusing effect is not applicable anymore, and the light might be collected by the output collimator directly. Consequently, the power limiting effect may not hold. However, this issue can be resolved by replacing the crossing-through prism with a total internal reflection structure, where the input beam is noncoaxial with the output. In this way, any damage to the material does not weaken the robustness of our proposal; instead, the device works as an optical fuse to permanently block the optical path.

IV. QUANTUM SIGNAL INTEGRITY

To determine if the proposed power limiter is useful for practical use, it is also important to study if the quantum signals can be disturbed when passing through the device. To this end, experiments based on time-bin (intensity), phase, and polarization encoding are implemented to see whether the quantum signal integrity can be affected.

Here, we study the quantum bit error rate (QBER) of the system, which is defined as the number of errors (N_{error}) over the total number of detection counts ($N_{\text{correct}} + N_{\text{error}}$),

$$\text{QBER} = \frac{N_{\text{error}}}{N_{\text{correct}} + N_{\text{error}}}. \tag{4}$$

In addition, given that the detector is well characterized (e.g., its background noise and single-photon efficiency are known), we may further write $\text{QBER} = \text{QBER}_{\text{opt}} + \text{QBER}_{\text{det}}$, where QBER_{opt} comes from quantum optical imperfections (e.g., imperfect state preparation, optical misalignment, etc.) and QBER_{det} comes from the detector dark counts. Here, as mentioned above, our main focus is the QBER_{opt} for intensity, phase, and polarization encoding schemes, which represent three of the most popular choices for QKD in practice.

The QBER of the intensity or time-bin encoding scheme is measured first. The intensity extinction ratio of a pulsed laser is measured to infer the QBER. The extinction ratio here is defined as the ratio between the count at the maximum state and the negative state. The pulsed laser is attenuated to about 0.1 photons per pulse and measured by an avalanche photodiode (APD) operating in the Geiger mode (gated). The laser pulse has a repetition frequency of 100 MHz and a pulse width of 400 ps. The APD has a gate width of 1 ns. The delay on the APD gate signal is scanned to cover both the laser pulse and dark region. The dark counts here are subtracted after the data acquisition for an accurate extinction ratio measurement of the optical pulse. The power limiter used here has a length of 101.6 mm and a diaphragm width of 750 μm. Average input powers of 14.49 and -19.72 dBm are tested to demonstrate the cases when the input power is close to and far below the power limiting threshold, respectively. The resulting counts as a function of delay are shown in Figs. 5(a) and 5(b). For the input power and the cases with and without the power limiter, the resulting extinction ratios are all above 35 dB, indicating a QBER of less than 0.032%. Therefore, we conclude that the introduction of the proposed power limiter does not introduce any significant noise to QKD systems based on time-bin encoding.

For the phase encoding scheme, the experimental setup is shown in Fig. 5(c). The input CW laser is modulated using a phase modulator switching between 0 and π phases with 50 MHz frequency. The laser output power is 10.28 dBm. The modulated signal is then decoded using an asymmetric Mach-Zehender interferometer (AMZI) with a path delay of around 10 ns. Moreover, a phase shifter is added in one of the paths of the AMZI to lock the relative phase. As such, the interference visibility, as well as the QBER, can be obtained. Finally, the output is attenuated to 0.1 photons per gate and measured by an APD. The counts as a function of delay are shown in Figs. 5(d) and 5(e), which correspond to the cases with and without power limiter installed, respectively. The interference visibility V is shown in Fig. 5(f) as $(1 - V)$ for a clear view. The maximum visibilities with and without the power limiter are 0.9844 and 0.9836, corresponding to QBERs of 0.78% and 0.82%, respectively. Thus, as in the case of time-bin encoding, we conclude that the proposed power limiter device is also suitable for phase-encoding QKD systems.

Finally, we study the impact of the device on polarization encoding. The experimental setup is shown in Fig. 5(g), where a CW laser with an output power of 11.41 dBm is used, and the polarization is manually tuned with a polarization controller. The attenuated output goes through a polarization beam splitter and the outputs are

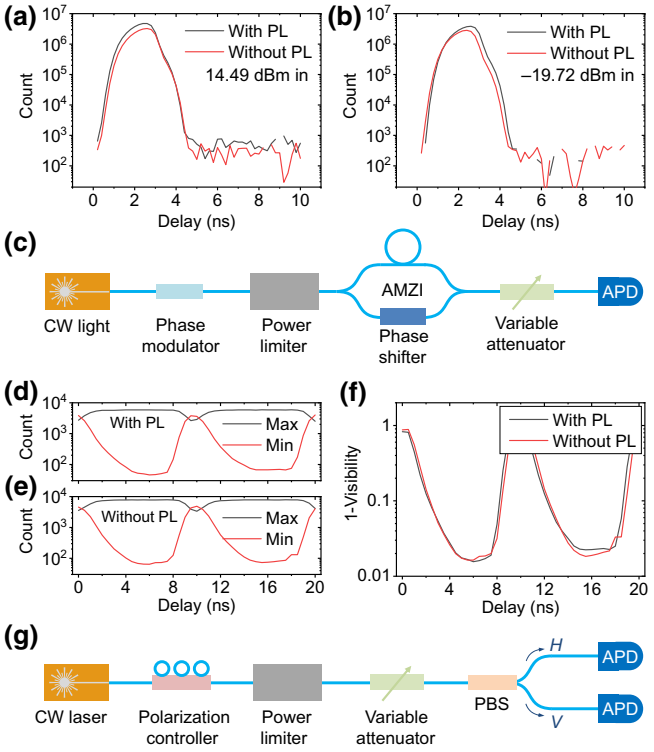

FIG. 5. (a),(b) The number of counts as a function of the delay between the APD gate and laser pulse in the intensity encoding scheme. The results with and without power limiter (PL) are shown. (a) The case for 14.49 dBm input power, which is close to the power limiting threshold. (b) The case for −19.72 dBm input power. The discontinuity of the curve is because of the negative count values obtained due to the statistical noise. (c) Experimental setup for the QBER measurement of the phase encoding scheme. The input CW light is modulated with a phase modulator switching between 0 and π phases and decoded with an AMZI, which generates a constructive or a destructive interference output, marked as max and min, respectively. The peaks at about 0 and 10 ns are caused by the AMZI path inaccuracy. The output light is attenuated to the single-photon level and measured with an APD. (d),(e) The count value with minimum and maximum phase shifter settings as a function of the delay between the APD gate and the phase modulator clock in the phase encoding scheme. (f) The calculated interference visibility as a function of the delay between the APD gate and the phase modulator clock with and without the power limiter in the phase encoding scheme. (g) Experimental setup for the QBER measurement of the polarization encoding scheme. The polarization of an attenuated CW laser is manually tuned to match the polarization of a polarization beam splitter (PBS). The output from the PBS is measured by two APDs.

measured by two APDs. The polarization extinction ratio is calculated from the ratio between the two APD counts. The result shows polarization extinction ratios of 30.1 and 32.6 dB for the cases with and without the power limiter, corresponding to QBERs of 0.098% and 0.055%, respectively. This clearly shows that the power limiter will not significantly disturb the state of polarization of the photon.

TABLE I. Summary for the extinction ratio (ER) or visibility (V) and QBER for different encoding schemes with and without PL.

Encoding schemes	With PL	ER (or V)	QBER (%)
Intensity (time bin)	Yes	> 35 dB	< 0.032
	No	> 35 dB	< 0.032
Phase	Yes	0.9844	0.78
	No	0.9836	0.82
Polarization	Yes	30.1 dB	0.098
	No	32.6 dB	0.055

All in all, we have experimentally confirmed that our power limiter device does not introduce any significant noise (in terms of the QBER) to standard QKD systems based on time, phase, and polarization encoding schemes. The results are also summarized in Table I. However, it should be noted that the power limiter does introduce extra losses (insertion loss) to the signal, so the photon collection efficiency would decrease when it is deployed on the receiver side. In our experiment, a minimum insertion loss of −5.1 dB is measured, which is equivalent to a transmission efficiency of around 31%, or a transmission distance of 25.5 km (assuming a single-mode fiber with a transmission loss of 0.2 dB/km). We note that this issue could be mitigated by using materials with higher TOC values, so that a smaller amount of light absorption is required to trigger the power limiting effect.

V. APPLICATIONS AND COUNTERMEASURES

A. Security against THAs

As an application of our proposed power limiter, we consider a phase-encoding coherent state MDI-QKD protocol [18,19,37] with energy-constrained THAs. A schematic of our system is shown in Fig. 6, where Alice and Bob are distant quantum transmitters and are supposed to prepare the required phase-encoding coherent states, then send them to Charlie for Bell-state measurement. The detailed protocol description can be found in Ref. [37]. A brief description of the protocol can also be found in Fig. 6.

To take THAs into consideration in security analyses, different models for Trojan-horse states have been proposed. For example, in Ref. [9], the Trojan-horse state is modeled as a pure coherent state with a fixed phase and intensity. However, this model might be too restrictive as Eve can send other states. In practice, she could send a mixture of coherent states with different intensities or other states that could potentially leak more information. Another model that can address potential THAs is presented in Refs. [38,39]. There, the nonvacuum component of the Trojan-horse state is modeled by an arbitrary

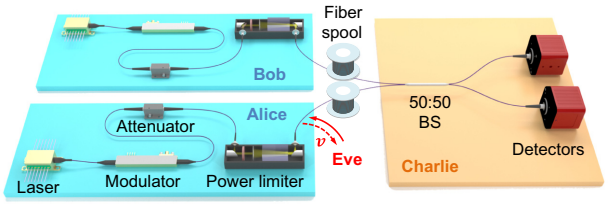

FIG. 6. Schematic of the phase-encoding MDI-QKD [37] system with the power limiter installed. For each round, using their lasers, modulators, and attenuators, Alice and Bob prepare one of the four coherent states $\{|\alpha e^{ix\pi/2}\rangle\}_x$ and $\{|\beta e^{iy\pi/2}\rangle\}_y$, where $x,y \in \{0,1,2,3\}$. They would then send their respective quantum states to Charlie for Bell-state measurement, who would then announce the outcome of his measurement $z \in \{L,R,\varnothing\}$. If Charlie announces $z=\varnothing$, they would discard the data for that round; otherwise, they would keep their respective choices x and y. After many rounds, they perform sifting and parameter estimation, and if they do not abort the protocol, they will also perform error correction and privacy amplification to extract identical and secret keys. Based on the power limiting threshold, the Trojan-horse attack from Eve could provide her with an average of ν Trojan-horse photons per pulse, which is taken into consideration for secure key rate calculation.

state that lives outside the qubit space in which the legitimate parties encode the information. While this model is very general and could take into account any source side channels, the resulting bounds can be overly pessimistic, since in the worst-case scenario the leakages might correspond to orthogonal quantum states and hence would leak full information about the modulation (key information).

In our analysis, we take the intermediate approach and allow Eve to send any Trojan-horse state in a given optical mode. On the other hand, we assume that the modulations are accurately characterized in our model. If the power limiting threshold is sufficiently low, the Trojan-horse state would have a large vacuum component that carries no information about the modulation. The crucial difference between our model and the model presented in Refs. [38,39] is that Eve needs to have multiphoton components to introduce orthogonality in the nonvacuum component of the Trojan horse, as we shall see in Eq. (6) below. This would in turn increase the vacuum contribution so that the average power remains below the threshold. On the other hand, since the leakage is completely uncharacterized in the framework of Refs. [38,39], in the worst-case scenario, a single photon might be sufficient to leak all the information. Therefore, we expect that our analysis would give better bounds against Trojan-horse attacks. Without loss of generality, the Trojan-horse state can be written as

$$|\xi\rangle = \sum_{n,m} c_{nm}|n\rangle|m\rangle|\mathcal{E}_{nm}\rangle, \tag{5}$$

where $|n\rangle, |m\rangle$ are the Fock states injected into Alice's and Bob's apparatus, respectively, and $|\mathcal{E}_{nm}\rangle$ is an ancilla that is kept in Eve's lab. The coefficients c_{nm} are the quantum amplitudes of the Fock states. Note that the state of the form (5) includes Trojan horses that are mixed (after tracing out Eve's ancilla) and may even be entangled.

The states $|n\rangle$ and $|m\rangle$ will accumulate some phases introduced by Alice's and Bob's modulators and hence they would leak some information about x and y. On the other hand, the state $|\mathcal{E}_{nm}\rangle$ will not accumulate any phase since it is kept in Eve's lab. After gathering the modulation information from the modulators, the output THA state thus has the form

$$|\xi'_{xy}\rangle = \sum_{n,m} c_{nm} e^{i(nx+my)\pi/2}|n\rangle|m\rangle|\mathcal{E}_{nm}\rangle. \tag{6}$$

Both the quantum states prepared by Alice and Bob and the THA state will be sent to Charlie via the quantum channel. Thus, the untrusted measurement can be modeled by a quantum-to-classical map, which can be described by an isometry $\mathcal{U}$ (with an appropriate purification):

$$\begin{aligned}|\phi_{xy}\rangle &= |e^{ix\pi/2}\alpha\rangle|e^{iy\pi/2}\beta\rangle|\xi'_{xy}\rangle \\ &\xrightarrow{\mathcal{U}} \sum_z |e^z_{xy}\rangle|z\rangle.\end{aligned} \tag{7}$$

Assuming that the effective states are of the form given in Eq. (7), the lower bound on the asymptotic secret key rate R is given by [40]

$$R \geq P_{\text{pass}}[1 - h_2(e_{\text{ph}}) - h_2(e_{\text{bit}})], \tag{8}$$

where p_{pass} is the probability of having a successful Bell-state measurement, $h_2(\cdot)$ is the binary entropy function, e_{bit} is the observed bit error rate, and e_{ph} is the phase-error rate that can be upper bounded using a simple modification of the technique presented in Refs. [41,42]. The detailed security analysis is presented in the Appendix. Importantly, our security analysis only requires the bound on the *average* power of the Trojan horses, which means that Eve is allowed to send a mixture of bright Trojan horse with vacuum.

To be more specific, based on the power limiting threshold obtained in Sec. III, a maximum photon number of injected eavesdropping light can be strictly constrained by the proposed optical power limiter. Then, the injected light will go through the attenuator twice before being collected by Eve, while the quantum state for QKD has just been attenuated once. Consider a QKD system working at a frequency of 1 GHz, with a power limiting threshold of 1 mW and an ideal phase modulator that does not introduce any extra insertion loss. In this case, an attenuation of 69 dB is sufficient to guarantee an average energy output of $\nu = 10^{-7}$. At the same time, the laser output can be adjusted to optimize the intensity μ for QKD, where

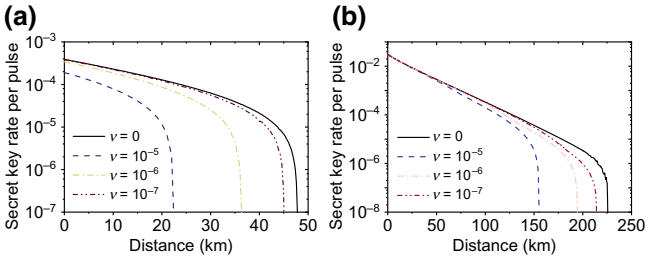

FIG. 7. Simulation of the asymptotic key rate for phase-encoding MDI QKD under two sets of parameters: (a) detector's efficiency $\eta_{\rm det} = 10\%$, dark count rate $p_{\rm dc} = 10^{-5}$, (b) detector's efficiency $\eta_{\rm det} = 85\%$, dark count rate $p_{\rm dc} = 10^{-7}$. Trojan-horse photon numbers of $\nu = 10^{-5}$, 10^{-6}, 10^{-7} and 0 are shown. The output intensity μ of each transmitter is optimized for each distance to maximize the key rate.

an averaged optical power of 23 μW can be used to generate quantum states with $\mu = 0.0183$. This is similar to the optimized intensity for MDI QKD with a detector efficiency of $\eta_{\rm det} = 85\%$, dark count rate $p_{\rm dc} = 10^{-7}$, and a 100 km transmission distance. Compared with Ref. [9] where the 12.8 W optical fiber damage threshold was used as the upper bound, the proposed power limiter could limit the power by 4 to 5 orders of magnitude lower. As a result, the requirement for attenuator and optical isolator is significantly reduced. Also, removing the need to have isolators could benefit future chip-based integration of such MDI-QKD systems.

To benchmark the performance of the protocol, we simulate the achievable asymptotic key rate with two sets of parameters: (1) detector's efficiency $\eta_{\rm det} = 10\%$, dark count rate $p_{\rm dc} = 10^{-5}$, (2) detector's efficiency $\eta_{\rm det} = 85\%$, dark count rate $p_{\rm dc} = 10^{-7}$. For both sets of parameters, the misalignment error $e_{\rm ali}$ is set to be 2%, and the transmission loss of the optical fiber is set to be 0.2 dB/km. We also assume that the central node is equidistant to Alice and Bob and $\mu_A = \mu_B = \mu$, which has been optimized over the simulation. As for the THA intensity, we set $\nu_A = \nu_B = \nu$. The results of the simulation are shown in Fig. 7. The results indicate that Alice and Bob can get a promising key rate without being affected much by the Trojan-horse attack if the energy of the THA is properly upper bounded.

B. Potential countermeasures

1. Bright illumination attacks

Laser damage attacks are a particularly powerful class of bright illumination attacks. In Refs. [11–13], it is shown that the detectors and optical components are prone to permanent changes and damages when Eve sends in a bright damaging laser with power of the order of watts. This is crucial because the security of most QKD systems depends on the integrity of their devices—that they behave according to design specifications.

Another class is detector blinding attacks [14–16]. By exploiting the implementation knowledge of single-photon detectors and the imperfect detector performance, Eve can send in a relatively strong eavesdropping light to change the working condition of the detector and get partial (or even full) control over the outcomes [14–16,44,45].

For illumination-related attacks, a common feature is that Eve must send in relatively bright light pulses. Hence, by restricting the input optical power using the proposed power limiter, it is expected that some of these attacks could be thwarted. To illustrate this possibility, we sketch out a method that could prevent the bright illumination attack presented in Ref. [43]; see Fig. 8(a). To start with, we note that standard single-photon detectors based on APD typically require a low-temperature operation to minimize the detectors' background noise, i.e., to limit the dark count rate. To cool the detectors, thermoelectric coolers are used, but these have limited cooling capacity. In Ref. [43], it is shown that injection of bright light pulses can create a situation in which the generated heat fails to dissipate completely. This leads to the breakdown voltage of the

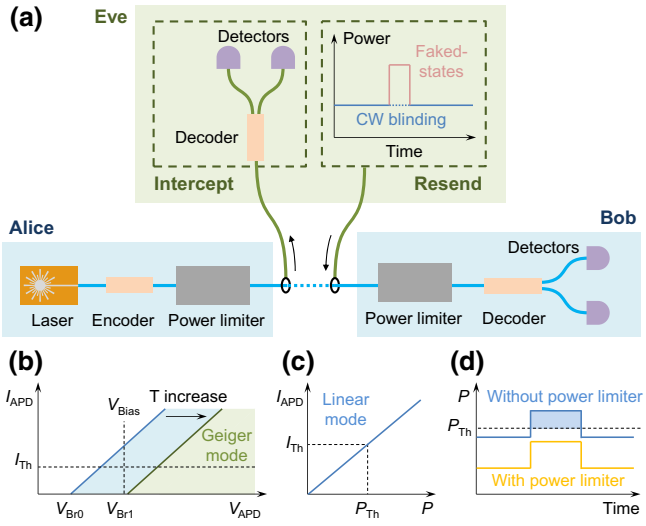

FIG. 8. (a) A schematic of the detector blinding attack proposed in Ref. [43]. (b) The current-voltage relationship of a typical APD. In normal working conditions, the breakdown voltage is $V_{\rm Br0}$. A fixed bias voltage $V_{\rm Bias} > V_{\rm Br0}$ enables Geiger mode operation with single-photon sensitivity. The resulting output current above a threshold $I_{\rm Th}$ is registered as a successful count. However, the breakdown voltage increases with the temperature T of the device. Thus, it is possible to blind the detector by driving it into the linear mode ($V_{\rm Bias} < V_{\rm Br1}$), thus nullifying the single-photon sensitivity. (c) The current-input power relationship of an APD operating in linear mode. By controlling the input power below or above the power threshold $P_{\rm Th}$, the detector could be controlled to register fake states. (d) The input power on Bob's detector with and without a power limiter.

APD going above the predetermined value, which consequently puts the detector into the *linear* mode (instead of the *Geiger* mode); see Fig. 8(b). In this case, the detector is no longer sensitive to single-photon input (i.e., *blinded*) and Eve can manipulate its outcome by sending a control light pulse superimposed on the bright light pulse, as depicted in Fig. 8(c).

According to Ref. [43], a bright CW light with an optical power of around 10 mW is required to blind the commercial QKD detectors. A control light pulse with a peak power of around 1 mW is sufficient to control the detector's outcome fully. Using a power limiter, as shown in Fig. 8(a), the input light power can be bounded below the blinding threshold, which would prevent the temperature of the detector from rising and hence from being blinded [see Fig. 8(d)]. For example, we can use an acrylic prism with a length of 50.8 mm and a diaphragm width of 380 μm to provide a power limiting threshold of 6.03 dBm (with -6.02 dB insertion loss) to prevent such attacks. It is important to note that, under normal working conditions, quantum signals (i.e., optical signals with small energy levels), in principle, only experience this small amount of loss when passing through the power limiter device. As such, the introduction of the proposed power limiter is unlikely to affect the overall performance of a QKD system. With our current design, we can expect a smaller insertion loss as well as a stronger power limiting effect, for example, by using a material with a higher TOC. However, it is important to mention that the proposed power limiter can only serve as a deterrence to detection-side-channel attacks based on bright illumination and not to all known attacks. Indeed, there are a variety of detection-side-channel attacks that do not require high power input, e.g., those based on detector superlinearity attacks [15,46], wavelength attacks [47], time-shifting attacks [48,49], and efficiency mismatch attacks [50]. To eliminate detection-side-channel attacks, the safest approach is to adopt MDI-QKD technology, as mentioned in the Introduction.

2. Plug-and-play QKD with untrusted light sources

Plug-and-play QKD is a two-way communication configuration [21] that aims to simplify implementation requirements such as polarization compensation and reference frame calibration. This approach is especially useful for practical MDI-QKD systems since it naturally guarantees near-perfect mode matching for the required two-photon interference [22,23,51].

By using external (untrusted) light sources instead of trusted light sources, however, plug-and-play systems are prone to transmitter-based attacks [6,52,53]. Again, the central issue here is that Eve can inject bright light pulses to break the working assumptions of QKD. To overcome this issue, one popular approach is to monitor the energy of the incoming light with a classical detector [22,54,55].

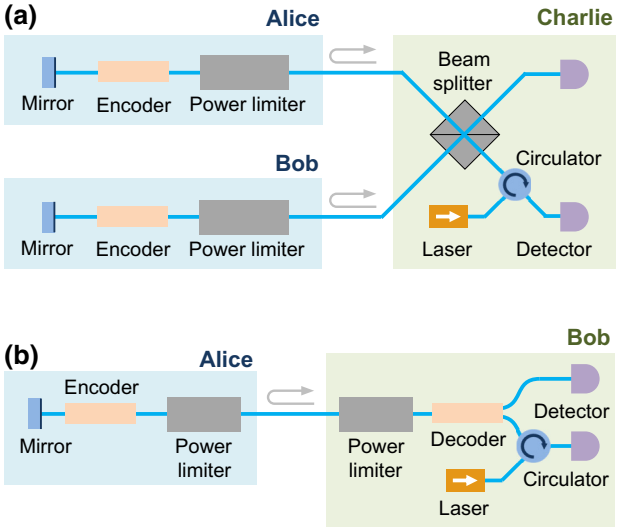

FIG. 9. Schematics of the power limiter used in (a) a plug-and-play MDI-QKD system and (b) a plug-and-play BB84 QKD system.

However, it has been shown that such active monitoring methods are not entirely robust and the classical detectors can still be hacked by exploiting their electrical circuitry; see, e.g., Refs. [12,56].

In light of the above, it is thus interesting to explore alternative countermeasures based on passive devices instead of active devices such as detectors. To this end, we propose to replace (or augment) the active power monitoring device with a passive power limiter, as shown in Fig. 9. Similar to the arguments provided in Sec. V A, the power limiter would limit the energy of the outgoing light, and hence Eve's knowledge about the key information as well; we leave a more careful security analysis to future work.

In addition, it is worthwhile to add that existing methods to limit incoming light energy are typically based on isolators, circulators and the laser damage threshold of devices [9,12,13]. These, however, are one-directional and add additional attenuation on the propagating direction of the eavesdropping light. In the case where the quantum signal has the same propagation as Eve's light, i.e., plug-and-play QKD or quantum receivers to resist against bright-illumination attacks, they may either pessimistically estimate Eve's information—since the actual input significantly deviates from the device damage threshold-, which will be used for security analysis, or introduce large insertion loss so the system performance will be greatly affected.

As a comparison, the proposed power limiter is shown to be able to provide an adjustable power limiting threshold on the output optical power and capable of protecting the system where the eavesdropping light and quantum signal have the same propagation direction.

VI. CONCLUSION

In this paper, we have proposed and demonstrated a passive power limiter design based on the thermo-optical defocusing effect of an acrylic prism. By numerical simulations and the experimental demonstration, we have rigorously studied the feasibility and performance of our power limiter design. In our experiment, the lowest optical power limiting threshold of −27.9 dBm with an insertion loss of −34.0 dB is measured. With a different setting, a low insertion loss of −5.1 dB is achieved with a 10.3 dBm power limiting threshold. The values are adjustable according to different system requirements. It is possible to further reduce the insertion loss at a certain power threshold by switching to a material with higher TOC values, for example thallium bromo-iodide KRS5 (-2.38×10^{-4} K^{-1}) [57] and silicone (-3.1×10^{-4} K^{-1}) [58], and/or reduce the beamwidth. Besides, our design possesses desirable features like compactness, robustness, plus polarization, and spectrum-dimension independence.

To illustrate the applicability of our proposed power limiter, we have quantitatively developed a general security analysis that allows for arbitrary Trojan-horse states. By properly limiting the THA energy leakage in an MDI-QKD system, a desirable secure key rate and transmission distance can be achieved. Moreover, based on previous evidence, we note that the power limiter can be useful for deterring bright illumination attacks in a quantum cryptography system. We take the thermal CW-blinding attack on the APD detectors as an example and show how the power limiter can be designed to prevent such an attack. We further discuss the possibility of using a power limiter to secure the plug-and-play QKD systems without active elements.

As demonstrated in our paper, by simply limiting the incoming or outgoing optical energy, a broad class of QKD protocols can be practically protected without introducing cumbersome device modification. Beyond these, one can also expect such a power limiting device to find applications in securing semi-device-independent quantum protocols based on energy or inner-product constraints [40,41,59–62], line-topology or ring-topology multiparty quantum communication systems [63,64], and long transmission distance QKD protocols such as the twin-field QKD [65–69]. As such, we believe that it will attract much interest and possess the potential to become a standard tool for quantum cryptography applications.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation (NRF) Singapore, under its NRF Fellowship programme (NRFF11-2019-0001) and Quantum Engineering Programme 1.0 projects (QEP-P2, QEP-P3, and QEP-P8)

APPENDIX: SECURITY ANALYSIS OF PHASE-ENCODING MDI QKD WITH TROJAN-HORSE ATTACK

In this appendix, we present the detailed security analysis of the phase-encoding MDI-QKD protocol presented in the main text. To analyze the security of the protocol, we modify the security proof technique presented in Ref. [41] to account for an arbitrary THA. Following the argument of Ref. [41], the untrusted measurement can be described by an isometry $\mathcal{U}$ acting on the effective signal state $|\phi_{xy}\rangle$ (which includes the Trojan-horse photons),

$$|\phi_{xy}\rangle \xrightarrow{\mathcal{U}} \sum_z |e^z_{xy}\rangle|z\rangle, \tag{A1}$$

where $|z\rangle$ is the classical register that stores the outcome of the Bell-state measurement and $|e^z_{xy}\rangle$ is a subnormalized state that stores Eve's quantum side information.

The Gram matrix associated with Eve's quantum side information, denoted by G, has to be positive semidefinite. Moreover, the observed statistics $\{P(z|x,y)\}_{x,y,z}$ put some constraints on G,

$$P(z|x,y) = \langle e^z_{xy}|e^z_{xy}\rangle. \tag{A2}$$

Furthermore, since the measurement is described by an isometry, the inner product has to be conserved, i.e.,

$$\langle\phi_{x'y'}|\phi_{xy}\rangle = \sum_z \langle e^z_{x'y'}|e^z_{xy}\rangle. \tag{A3}$$

Finally, the phase error rate, which quantifies how much information is leaked to Eve, is a linear function of the Gram matrix. Following the derivation in Appendix A of Ref. [41], the phase error rate is given by

$$\begin{aligned} e_{\text{ph}} = \frac{1}{2} + \frac{1}{4P_{\text{pass}}}\text{Re}\{&\langle e^L_{00}|e^L_{22}\rangle - \langle e^R_{00}|e^R_{22}\rangle \\ &- \langle e^L_{02}|e^L_{20}\rangle + \langle e^R_{02}|e^R_{20}\rangle\}, \end{aligned} \tag{A4}$$

where

$$\begin{aligned} P_{\text{pass}} = \tfrac{1}{4}[&P(L|00) + P(R|00) + P(L|02) + P(R|02) \\ &+ P(L|20) + P(R|20) + P(L|22) + P(R|22)] \end{aligned} \tag{A5}$$

is the probability of a successful Bell-state measurement given that Alice and Bob choose the key generation basis (i.e., when $x, y \in \{0, 2\}$), which is observed directly in the experiment. The linearity of the constraints and the objective function allows us to use semidefinite programming to find a tight bound on the phase error rate e_{ph}.

For fixed observed statistics, the phase error rate e_{ph} depends only on the Gram matrix of the effective signal

states $\{|\phi_{xy}\rangle\}_{x,y}$. Our goal is therefore to characterize the set of Gram matrices of the effective signal states subject to the constraint on the mean energy of the Trojan-horse lights. Indeed, the main difference between our security analysis and that presented in Ref. [41] lies in the fact that $\langle\phi_{x'y'}|\phi_{xy}\rangle$ is not perfectly characterized, but it can be bounded due to the energy constraints on the Trojan-horse state.

Assuming that Alice and Bob each modulate lights from a single mode, the most general form of Trojan-horse state that Eve can send is given by

$$|\xi\rangle_{abe} = \sum_{n,m} c_{nm}|n\rangle_a|m\rangle_b|\mathcal{E}_{nm}\rangle_e, \tag{A6}$$

where the registers a, b, e are held in Alice's, Bob's, and Eve's labs, respectively. The states $|n\rangle$ and $|m\rangle$ denote photon number states. In general, the Trojan horse can also be entangled to some ancillary system e that is kept in Eve's lab, which we denote by $|\mathcal{E}_{nm}\rangle$. In practice, only a fraction of the Trojan-horse lights is reflected out to the quantum channel. The rest of the photons are lost (and, therefore, are inaccessible to Eve). In our model, we conservatively ignore the loss in the modulators. Finally, for the remainder of this section, we omit the subscripts denoting the registers when there is no danger of ambiguity.

Now, Alice's and Bob's phase modulations can be described by the unitary operators

$$\hat{U}_x = e^{i\pi x\hat{n}/2}, \qquad \hat{U}_y = e^{i\pi y\hat{m}/2}, \tag{A7}$$

where $\hat{n}$ and $\hat{m}$ are the number operators acting on the a and b registers, respectively. Thus, after the phase modulations, the Trojan horse evolves into

$$|\xi'_{xy}\rangle = \hat{U}_x \otimes \hat{U}_y \otimes \mathbb{1}|\xi\rangle = \sum_{n,m} c_{nm}e^{i(nx+my)\pi/2}|n\rangle|m\rangle|\mathcal{E}_{nm}\rangle, \tag{A8}$$

which gives the effective state

$$|\phi_{xy}\rangle = |e^{ix\pi/2}\alpha\rangle|e^{iy\pi/2}\beta\rangle|\xi'_{xy}\rangle, \tag{A9}$$

when Alice is given the input x and Bob is given the input y. Hence, for a fixed combination x, y, x', y', we have

$$\langle\phi_{x'y'}|\phi_{xy}\rangle = \left(\sum_{nm}|c_{nm}|^2 e^{i[n(x-x')+m(y-y')]\pi/2}\right) \times \langle e^{ix'\pi/2}\alpha|e^{ix\pi/2}\alpha\rangle\langle e^{iy'\pi/2}\beta|e^{iy\pi/2}\beta\rangle, \tag{A10}$$

where the term in the parentheses is the contribution due to the THA. Here, α and β denote the amplitudes of Alice's and Bob's lasers.

Now, due to the symmetry of the phase modulations, observe that

$$\begin{aligned} e^{i(n+4)x\pi/2} &= e^{inx\pi/2}e^{i2\pi x} = e^{inx\pi/2}, \\ e^{i(m+4)y\pi/2} &= e^{imy\pi/2}e^{i2\pi y} = e^{imy\pi/2}, \end{aligned} \tag{A11}$$

for all $x, y \in \{0, 1, 2, 3\}$. Hence, without loss of generality, it is sufficient to consider $n, m \in \{0, 1, 2, 3\}$. Denoting the probability that Alice and Bob respectively receive n and m photons by $|c_{nm}|^2 = P_{nm}$, it is therefore sufficient to consider a finite number of $\{P_{nm}\}_{n,m}$.

Here, in contrast to the analysis presented in Ref. [41], the Gram matrix of the effective signal states are not fixed as Eve can vary the photon number distribution to maximize the leakage. Therefore, there are two variables that we consider in our optimization, namely the Gram matrix of Eve's quantum side information (which we denote by G) and the photon number distribution (which we denote by P_{nm}). Therefore, taking into account Eve's freedom to choose the Trojan-horse state, we have to solve the optimization problem

$$\begin{aligned} \max_{G,P_{nm}} \quad & e_{\mathrm{ph}} \\ \text{subject to} \quad & G \succeq 0, \\ & e_{\mathrm{ph}} \le 1/2, \\ & P(z|x,y) = \langle e^z_{xy}|e^z_{xy}\rangle, \\ & P_{nm} \ge 0, \\ & \sum_{n,m} P_{nm} = 1, \\ & \sum_{n,m} P_{nm}n \le \nu_A, \\ & \sum_{n,m} P_{nm}m \le \nu_B, \\ & \sum_z \langle e^z_{x'y'}|e^z_{xy}\rangle = \sum_{nm} P_{nm}e^{i[n(x-x')+m(y-y')]\pi/2}\Lambda_{x'y',xy}, \end{aligned} \tag{A12}$$

where $n, m \in \{0, 1, 2, 3\}$, ν_A and ν_B denote the intensity of the Trojan-horse lights (measured at the outputs of Alice's and Bob's sources, respectively), and

$$\Lambda_{x'y',xy} = \langle e^{ix'\pi/2}\alpha|e^{ix\pi/2}\alpha\rangle\langle e^{iy'\pi/2}\beta|e^{iy\pi/2}\beta\rangle \tag{A13}$$

is the inner product of Alice's and Bob's characterized signal states (i.e., the signal states in the absence of the THA).

One could then substitute the bound on e_{ph} into the key rate formula

$$R \geq P_{\mathrm{pass}}[1 - h_2(e_{\mathrm{ph}}) - h_2(e_{\mathrm{bit}})], \tag{A14}$$

where $h_2(\cdot)$ is the binary entropy function and e_{bit} is the bit error rate in the key generation basis.

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

[2] E. Diamanti, H.-K. Lo, B. Qi, and Z. Yuan, Practical challenges in quantum key distribution, npj Quantum Inf. **2**, 16025 (2016).

[3] 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).

[4] 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).

[5] A. Vakhitov, V. Makarov, and D. R. Hjelme, Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography, J. Mod. Opt. **48**, 2023 (2001).

[6] N. Gisin, S. Fasel, B. Kraus, H. Zbinden, and G. Ribordy, Trojan-horse attacks on quantum-key-distribution systems, Phys. Rev. A **73**, 022320 (2006).

[7] N. Jain, E. Anisimova, I. Khan, V. Makarov, C. Marquardt, and G. Leuchs, Trojan-horse attacks threaten the security of practical quantum cryptography, New J. Phys. **16**, 123030 (2014).

[8] S. Sajeed, C. Minshull, N. Jain, and V. Makarov, Invisible Trojan-horse attack, Sci. Rep. **7**, 8403 (2017).

[9] 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).

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

[11] A. N. Bugge, S. Sauge, A. M. M. Ghazali, J. Skaar, L. Lydersen, and V. Makarov, Laser Damage Helps the Eavesdropper in Quantum Cryptography, Phys. Rev. Lett. **112**, 070503 (2014).

[12] V. Makarov, J.-P. Bourgoin, P. Chaiwongkhot, M. Gagné, T. Jennewein, S. Kaiser, R. Kashyap, M. Legré, C. Minshull, and S. Sajeed, Creation of backdoors in quantum communications via laser damage, Phys. Rev. A **94**, 030302 (2016).

[13] A. Huang, R. Li, V. Egorov, S. Tchouragoulov, K. Kumar, and V. Makarov, Laser-Damage Attack Against Optical Attenuators in Quantum Key Distribution, Phys. Rev. Appl. **13**, 034017 (2020).

[14] V. Makarov, Controlling passively quenched single photon detectors by bright light, New J. Phys. **11**, 065003 (2009).

[15] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Hacking commercial quantum cryptography systems by tailored bright illumination, Nat. Photonics **4**, 686 (2010).

[16] Z. L. Yuan, J. F. Dynes, and A. J. Shields, Resilience of gated avalanche photodiodes against bright illumination attacks in quantum cryptography, Appl. Phys. Lett. **98**, 231104 (2011).

[17] Z. L. Yuan, J. F. Dynes, and A. J. Shields, Avoiding the blinding attack in QKD, Nat. Photonics **4**, 800 (2010).

[18] H.-K. Lo, M. Curty, and B. Qi, Measurement-Device-Independent Quantum Key Distribution, Phys. Rev. Lett. **108**, 130503 (2012).

[19] S. L. Braunstein and S. Pirandola, Side-Channel-Free Quantum Key Distribution, Phys. Rev. Lett. **108**, 130502 (2012).

[20] C. Wang, G. Zhang, and C. C. W. Lim, Method and device for optical power limiter, SG Non-Provisional Application No. 10202006635S.

[21] A. Muller, T. Herzog, B. Huttner, W. Tittel, H. Zbinden, and N. Gisin, "Plug and play" systems for quantum cryptography, Appl. Phys. Lett. **70**, 793 (1997).

[22] F. Xu, Measurement-device-independent quantum communication with an untrusted source, Phys. Rev. A **92**, 012333 (2015).

[23] G.-Z. Tang, S.-H. Sun, F. Xu, H. Chen, C.-Y. Li, and L.-M. Liang, Experimental asymmetric plug-and-play measurement-device-independent quantum key distribution, Phys. Rev. A **94**, 032326 (2016).

[24] D. C. Smith, High-power laser propagation: Thermal blooming, Proc. IEEE **65**, 1679 (1977).

[25] R. C. C. Leite, S. P. S. Porto, and T. C. Damen, The thermal lens effect as a power-limiting device, Appl. Phys. Lett. **10**, 100 (1967).

[26] M. E. DeRosa and S. L. Logunov, Fiber-optic power limiter based on photothermal defocusing in an optical polymer, Appl. Opt. **42**, 2683 (2003).

[27] D. Smith, Thermal defocusing of CO_2 laser radiation in gases, IEEE J. Quantum Electron. **5**, 600 (1969).

[28] Z. Zhang, P. Zhao, P. Lin, and F. Sun, Thermo-optic coefficients of polymers for optical waveguide applications, Polymer **47**, 4893 (2006).

[29] X. Zhang, J. Qiu, X. Li, J. Zhao, and L. Liu, Complex refractive indices measurements of polymers in visible and near-infrared bands, Appl. Opt. **59**, 2337 (2020).

[30] S. Rudtsch and U. Hammerschmidt, Intercomparison of measurements of the thermophysical properties of polymethyl methacrylate, Int. J. Thermophys. **25**, 1475 (2004).

[31] P. A. Soave, R. A. F. Dau, M. R. Becker, M. B. Pereira, and F. Horowitz, Refractive index control in bicomponent polymer films for integrated thermo-optical applications, Opt. Eng. **48**, 124603 (2009).

[32] F. Qiu, D. Yang, G. Cao, R. Zhang, and P. Li, Synthesis, characterization, thermal stability and thermo-optical properties of poly(urethane-imide), Sensors Actuators B: Chem. **135**, 449 (2009).

[33] G. P. Agrawal, in *Nonlinear Science at the Dawn of the 21st Century* (Springer-Verlag, Berlin Heidelberg, Germany, 2000), p. 195.

[34] X. Zhang, J. Qiu, J. Zhao, X. Li, and L. Liu, Complex refractive indices measurements of polymers in infrared bands, Quant. Spectrosc. Radiat. Transfer **252**, 107063 (2020).

[35] R. M. Taha, Hole drilling in polymethyl methacrylate (PMMA) using CO_2 laser, Diyala J. Eng. Sci. **7**, 30 (2014).

[36] P. Berrie and F. Birkett, The drilling and cutting of polymethyl methacrylate (perspex) by CO_2 laser, Opt. Lasers Eng. **1**, 107 (1980).

[37] K. Tamaki, H.-K. Lo, C.-H. F. Fung, and B. Qi, Phase encoding schemes for measurement-device-independent quantum key distribution with basis-dependent flaw, Phys. Rev. A **85**, 042307 (2012).

[38] M. Pereira, M. Curty, and K. Tamaki, Quantum key distribution with flawed and leaky sources, npj Quantum Inf. **5**, 1 (2019).

[39] M. Pereira, G. Kato, A. Mizutani, M. Curty, and K. Tamaki, Quantum key distribution with correlated sources, Sci. Adv. **6**, eaaz4487 (2020).

[40] P. W. Shor and J. Preskill, Simple Proof of Security of the BB84 Quantum Key Distribution Protocol, Phys. Rev. Lett. **85**, 441 (2000).

[41] Y. Wang, I. W. Primaatmaja, E. Lavie, A. Varvitsiotis, and C. C. W. Lim, Characterising the correlations of prepare-and-measure quantum networks, npj Quantum Inf. **5**, 1 (2019).

[42] I. W. Primaatmaja, E. Lavie, K. T. Goh, C. Wang, and C. C. W. Lim, Versatile security analysis of measurement-device-independent quantum key distribution, Phys. Rev. A **99**, 062332 (2019).

[43] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Thermal blinding of gated detectors in quantum cryptography, Opt. Express **18**, 27938 (2010).

[44] L. Lydersen, M. K. Akhlaghi, A. H. Majedi, J. Skaar, and V. Makarov, Controlling a superconducting nanowire single-photon detector using tailored bright illumination, New J. Phys. **13**, 113042 (2011).

[45] H. Qin, R. Kumar, V. Makarov, and R. Alléaume, Homodyne-detector-blinding attack in continuous-variable quantum key distribution, Phys. Rev. A **98**, 012312 (2018).

[46] L. Lydersen, N. Jain, C. Wittmann, Ø. Marøy, J. Skaar, C. Marquardt, V. Makarov, and G. Leuchs, Superlinear threshold detectors in quantum cryptography, Phys. Rev. A **84**, 032320 (2011).

[47] J.-Z. Huang, C. Weedbrook, Z.-Q. Yin, S. Wang, H.-W. Li, W. Chen, G.-C. Guo, and Z.-F. Han, Quantum hacking of a continuous-variable quantum-key-distribution system using a wavelength attack, Phys. Rev. A **87**, 062329 (2013).

[48] B. Qi, C.-H. F. Fung, H.-K. Lo, and X.-F. Ma, Time-shift attack in practical quantum cryptosystems, Quantum Inf. Comput. **7**, 73 (2007).

[49] Y. Zhao, C.-H. F. Fung, B. Qi, C. Chen, and H.-K. Lo, Quantum hacking: Experimental demonstration of time-shift attack against practical quantum-key-distribution systems, Phys. Rev. A **78**, 042333 (2008).

[50] V. Makarov, A. Anisimov, and J. Skaar, Effects of detector efficiency mismatch on security of quantum cryptosystems, Phys. Rev. A **74**, 022313 (2006).

[51] C. Wang, X.-T. Song, Z.-Q. Yin, S. Wang, W. Chen, C.-M. Zhang, G.-C. Guo, and Z.-F. Han, Phase-Reference-Free Experiment of Measurement-Device-Independent Quantum key Distribution, Phys. Rev. Lett. **115**, 160502 (2015).

[52] F. Xu, B. Qi, and H.-K. Lo, Experimental demonstration of phase-remapping attack in a practical quantum key distribution system, New J. Phys. **12**, 113026 (2010).

[53] S.-H. Sun, M.-S. Jiang, and L.-M. Liang, Passive Faraday-mirror attack in a practical two-way quantum-key-distribution system, Phys. Rev. A **83**, 062331 (2011).

[54] Y. Zhao, B. Qi, and H.-K. Lo, Quantum key distribution with an unknown and untrusted source, Phys. Rev. A **77**, 052327 (2008).

[55] Y. Zhao, B. Qi, H.-K. Lo, and L. Qian, Security analysis of an untrusted source for quantum key distribution: Passive approach, New J. Phys. **12**, 023024 (2010).

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

[57] G. Ghosh, *Handbook of Optical Constants of Solids: Handbook of Thermo-Optic Coefficients of Optical Materials with Applications* (Academic Press, Cambridge, Massachusetts, 1998).

[58] Z. Zhang, P. Zhao, P. Lin, and F. Sun, Thermo-optic coefficients of polymers for optical waveguide applications, Polymer **47**, 4893 (2006).

[59] T. Van Himbeeck, E. Woodhead, N. J. Cerf, R. Garcia-Patron, and S. Pironio, Semi-device-independent framework based on natural physical assumptions, Quantum **1**, 33 (2017).

[60] T. Van Himbeeck and S. Pironio, Correlations and randomness generation based on energy constraints, ArXiv:1905.09117 (2019).

[61] M. Avesani, H. Tebyanian, P. Villoresi, and G. Vallone, Semi-device-independent heterodyne-based quantum random number generator, ArXiv:2004.08344 (2020).

[62] D. Rusca, T. van Himbeeck, A. Martin, J. B. Brask, W. Shi, S. Pironio, N. Brunner, and H. Zbinden, Self-testing quantum random-number generator based on an energy bound, Phys. Rev. A **100**, 062338 (2019).

[63] W. P. Grice, P. G. Evans, B. Lawrie, M. Legré, P. Lougovski, W. Ray, B. P. Williams, B. Qi, and A. M. Smith, Two-party secret key distribution via a modified quantum secret sharing protocol, Opt. Express **23**, 7300 (2015).

[64] C. Schmid, P. Trojek, M. Bourennane, C. Kurtsiefer, M. Żukowski, and H. Weinfurter, Experimental Single Qubit Quantum Secret Sharing, Phys. Rev. Lett. **95**, 230505 (2005).

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

[66] J.-P. Chen, C. Zhang, Y. Liu, C. Jiang, W. Zhang, X.-L. Hu, J.-Y. Guan, Z.-W. Yu, H. Xu, J. Lin, *et al.*, Sending-Or-Not-Sending With Independent Lasers: Secure Twin-Field Quantum Key Distribution Over 509 km, Phys. Rev. Lett. **124**, 070501 (2020).

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

[68] S. Wang, D.-Y. He, Z.-Q. Yin, F.-Y. Lu, C.-H. Cui, W. Chen, Z. Zhou, G.-C. Guo, and Z.-F. Han, Beating the Fundamental Rate-Distance Limit in a Proof-Of-Principle Quantum Key Distribution System, Phys. Rev. X **9**, 021046 (2019).

[69] X.-Q. Zhong, J.-Y. Hu, M. Curty, L. Qian, and H.-K. Lo, Proof-Of-Principle Experimental Demonstration of Twin-Field Type Quantum Key Distribution, Phys. Rev. Lett. **123**, 100506 (2019).