Quantum key distribution protocols [1, 2] harness superposition and entanglement to ensure secure information transfer between two parties [3]. However, the long-standing bottleneck of long-distance quantum communication are photonic losses. The quantum repeater (QR) was proposed in Ref. [4] as a general solution to this problem by dividing a long-distance link into segments, and making use of entanglement swapping between the nodes to generate entanglement over the long-distance link. An essential requirement for QRs for ground- and satellite-based quantum networks [5–7] are quantum memories (QM) [8, 9]. These act as interfaces between flying and stationary qubits and allow for storage of quantum information, ideally in an efficient and noise-free manner. QMs are being developed in many different platforms, ranging from ultracold atoms [10–12] to solid-state systems [13–17] and warm atomic vapors [18–22]. The latter constitute technologically simple and scalable systems, more suitable for use on satellites, as neither laser cooling, strong magnetic fields nor ultra-high vacuum are needed for their implementation. Due to a high optical depth, high efficiencies are possible. However, these types of memories are often considered to be prone to noise, especially four-wave-mixing (FWM) [23].

Building secure global quantum communication networks [24] will make it necessary to pair such memories with bright and efficient single-photon sources such as semiconductor quantum dots (QDs) [25–29], spontaneous parametric down-conversion (SPDC) sources [30–33], or sources based on room temperature atomic ensembles [34, 35]. Essential requirements on compatible quantum memories for this endeavor are high end-to-end efficiencies $\eta_{\text{e2e}}$, high signal-to-noise ratios (SNR), and sufficiently long storage times in the millisecond range. In order to benchmark the memory performance, we use the parameter $\mu_1$, defined as the mean photon number in the input that results in a SNR of 1 in the output [36]. In Refs. [5, 6], key rates for specific QR configurations are simulated with realistic memory parameters, such as memory efficiencies of $\eta_{\text{mem}} = 70 - 80\%$ and storage times of some milliseconds. Previously, an internal memory efficiency of $\eta_{\text{mem}} = 80\%$, without considering the filtering system [21], and a low noise figure of $\mu_1 = 0.20(2)$ [22] have been reported. A $1/e$ storage time of $t_{\text{store}} = 430$ ms for attenuated coherent pulses was also achieved [20]. Nevertheless, in these state-of-the-art implementations only one parameter ($\eta_{\text{e2e}}, \mu_1, t_{\text{store}}$) was optimized at a time. Especially regarding applications in quantum communication, it is imperative to find operating conditions which maximize all relevant parameters simultaneously.

Our work focuses on the simultaneous optimization of $\eta_{\text{e2e}}$ and $\mu_1$. The used EIT $\Lambda$-configuration with co-propagating signal and control is in principle suitable for long storage times above 1 ms, which will be pursued in follow-up experiments. We achieve $\eta_{\text{e2e}} = 13(2)\%$ and $\mu_1 = 0.07(2)$ simultaneously. The internal memory efficiency reaches $\eta_{\text{mem}} = 33(1)\%$. The origin of the readout noise is found to be a combination of spontaneous Raman scattering (SRS) and fluorescence noise, where the former dominates at larger control pulse energies. Apart from practical applications in long-distance QKD, these results are of general interest for the community involved in the exploitation of laser-induced atomic coherence. The presented hyperfine ground-state memory scheme is based on EIT [37, 38] on the Cs D1 transition, following the scheme shown in Fig.1(a), and is, to our knowledge, the first realization of a warm single-photon-level memory in cesium on this transition, although there exist realizations using cold atoms, e.g. Ref. [39]. Cesium-
fers a higher possible bandwidth, due to a larger ground-state hyperfine splitting, as compared to e.g. rubidium. Comparatively, cesium also shows a larger excited state hyperfine splitting, which reduces the influence of the additional, unwanted excited state. Further advantages of cesium include its larger mass and longer D1 wavelength, which result in slower thermal movement and therefore less Doppler broadening, and in less absorption in optical fibers compared to the Rb D1 line, respectively. As depicted in the experimental setup in Fig. 1(b), the orthogonal and linearly polarized signal (S) and control (C) lasers (ECDLs from Sacher Lasertechnik) are red-detuned by Δ from the F = 3 → F′ = 3 and F = 4 → F′ = 3 transitions respectively, and are offset-locked to a frequency difference of ~ 9.2 GHz. A small deviation from the exact hyperfine splitting is caused by the AC-Stark shift induced by the control laser. Gaussian signal and control pulses of varying full width at half maximum (FWHM) are generated by fiber-based electro-optic modulators controlled by arbitrary function generators. The control pulses are amplified by a semiconductor optical amplifier (SOA). The SOAs amplified spontaneous emission is suppressed by a combination of a narrow-band dielectric filter (1 nm FWHM) and a Fabry-Pérot etalon (free spectral range FSR = 205.5 GHz, finesse F = 47). This results in a control laser peak power of P_{C,max} = 12.9 mW. For single-photon-level measurements the signal pulses are attenuated with optical filters. To calibrate the photon number, 10% of the signal pulse (N_{mon}) is monitored on an avalanche photodiode (APD). The signal and control beams are overlapped using a calcite prism and sent through the Cs memory cell. Hereby, we put special care into making signal and control beams co-propagate through the cylindrical vapor cell (7.5 cm length, 2 cm diameter). The cell contains $^{133}$Cs and 5 Torr N₂ buffer gas, is enclosed inside a µ-metal magnetic shielding to nominally reduce the oscillating magnetic field in its interior by a factor of 1000, and is kept at T_{cell} = 60°C. We estimate the wavelength of the spin wave due to the residual angle to be $\lambda_{sw} = 2\pi/|\vec{k}_{S} - \vec{k}_{C}| \approx 5$ mm. The control and signal beam diameters are 109(5) µm and 93(5) µm FWHM at their focus at the center of the memory cell, and their Rayleigh length (as for all beams) extends until the cell’s end. With a dipole moment of $\alpha = 2.7 \cdot 10^{-29}$ C·m [40] for the Cs D1 transition, this yields a peak control Rabi frequency of $\Omega_{C} \approx 2\pi \cdot 540$ MHz. A pump laser on the F = 4 → F′ = 4 transition, turned on and off using an acousto-optic modulator (AOM), is used for state preparation in the F = 3 ground state ($|g\rangle$). Its beam diameter is 540(5) µm FWHM at its focus, with a peak power of $P_{pump} = 12.3$ mW. The measured pumping efficiency at this configuration is approximately $n_{F=3}/n_{F=4} = 80\%$. To filter the signal from the control pulses we employ a combination of polarization- and spectral filtering. The former yields a suppression by 7 orders of
magnitude for the control light. The following spectral filtering is accomplished by two Fabry-Pérot etalons (FSR = 25.6 GHz, $f = 47$), each of which further attenuates the control laser by 3 orders of magnitude within the signal beam path. In total, this amounts to a control laser suppression by 13 orders of magnitude. Meanwhile, the signal light is only attenuated by a factor of 2.5. At the output, the signal photons are detected by a single-photon-counting APD (Laser Components COUNT-50C-FC) and the resulting signal is analyzed using time-correlated single photon counting.

A single run of the storage-retrieval experiment begins by switching on the pump laser for 10 μs to prepare the state $|g⟩$. After pumping, the control and signal pulses enter the memory. Hereby, the signal pulse is delayed for several ns with respect to the control pulse, as depicted in Fig. 1(c). To retrieve the signal photon from the memory, a second identical control pulse is sent into the cell after $t = 200$ ns. The storage and retrieval experiment is repeated at a rate $f_{\text{rep}} = 1/(11\mu s)$ for an integration time $t_{\text{int}}$ of 60 s.

Figure 1(d) shows a typical measured photon arrival time histogram as used for all further analysis. Shown are measured data for storage and retrieval of a coherent state containing $|α|^2 = (N_{\text{mon}} σ)/(f_{\text{rep}} η_{\text{APD}}) = 1.0(1)$ photons on average, a noise measurement resulting from storage of a vacuum state with $|α|^2 = 0$, arising from a blocked input, and the noise-corrected signal. The uncertainty of $|α|^2$ arises from the error in the signal/monitor splitting ratio $σ$, measured with a power-meter (Thorlabs PM1160); the uncertainty of the APD efficiency $η_{\text{APD}} = 0.33(5)$, and the statistical error in the monitored photon number $N_{\text{mon}}$. The peak at $t = 0$ ns corresponds to the part of the signal pulse that is transmitted through the memory (leakage). The peak at $t = 140$ ns corresponds to the signal read out from the memory. As shown in Figure 1(d), the temporal profile of the retrieved signal peak is distorted by the memory and shows two distinct maxima. We define the storage time $t_{\text{storage}}$ as the difference between the first (and highest) of these maxima, and the peak of the incoming signal at $t = 0$ ns. The used storage times were chosen to be short in order to focus on the $η_{2e}$ and SNR, but can be extended into the μs range with the present setup, and into the millisecond range by increasing the beam diameters and changing the Λ-system configuration by exploiting the magnetic Zeeman sublevels of the hyperfine ground states [20].

Given the temporal envelope of the retrieved signal, the $η_{2e}$ of the memory setup, including the filtering system, and the SNR at single photon level are given as

$$η_{2e} = \frac{N_{\text{signal}} - N_{\text{noise}}}{|α|^2 η_{\text{APD}} f_{\text{rep}} t_{\text{int}}},$$

$$\text{SNR}_α = \frac{N_{\text{signal}, α} - N_{\text{noise}}}{N_{\text{noise}}},$$

where $N_{\text{signal}, α}$ ($N_{\text{noise}}$) correspond to the number of signal(noise) counts. Due to the long integration time, the statistical error is small and the uncertainty in $η_{2e}$ is dominated by the systematic error in the power measurement and the calibration of the single photon level. For the SNR, as it is measured at single-photon level, the error bars are proportional to the uncertainty of $|α|^2$. The $η_{2e}$ increases with the length of the retrieval time integration window $t_{\text{max}}$, while the SNR decreases, as can be taken from the inset in Fig.1(d). Hence, a compromising value of $t_{\text{max}} = 155$ ns was chosen. The lower integration limit is always taken at the minimum before each retrieval peak. Figure 2 shows the dependence of $η_{2e}$ and SNR on signal pulse width $Δt_S$ (a), energy in the control pulses $E_C$ (b), and detuning $Δ$ (c). $Δt_S$ is related to the bandwidth $Δω_S$ (FWHM) of the pulses that are stored in the memory as $Δt_S Δω_S = 2 ln(2)/π$, for transform-limited pulses. For the measurements we took $Δt_S = 25$ ns and $Δ = 2300(100)$ MHz as baseline and varied only the respective parameter. $E_C$ was hereby adjusted in a range $280(50)$ pJ $\leq E_C \leq 560(50)$ pJ. In all three cases, $η_{2e}$ shows a saturation behavior, while the SNR also saturates in Fig. 2(a), but reaches a maximum in Figs. 2(b) and (c), and then decreases again. Therefore, there exists a trade-off between reaching the maximally possible efficiency, and adding more noise to the retrieval signal. We note that the large scatter in the SNR visible in Fig. 2(c) was due to imperfect measurement in the early stage of the experiment. With more accurate measurements performed later on, monitoring the laser frequency using a wavemeter (Wavelength Meter WS/7-60), such a large scatter was not visible anymore. To describe $η_{2e}$, the following models were used: The efficiency as a function of signal pulse width (Fig. 2(a)) is $η_{2e}(Δt_S) = η_{0,τ_S}/\sqrt{1+(4ln(2) Δω_{\text{mem}})^2}$, where $η_{0,τ_S}$ is the maximal achievable efficiency and $Δω_{\text{mem}}$ is the bandwidth of the memory. This model arises from the Maxwell-Bloch equations and describes how efficiently a pulse of given width $Δt_S$ is converted into an atomic spin wave, when limited by the available control Rabi frequency $Ω_C$, which determines the spectral width of the EIT window, and therefore the memory bandwidth [41]. It is in good agreement with the measured data and yields $η_{0,τ_S} = 0.128(7)$ and $Δω_{\text{mem}} = 220(30)$ MHz. In Fig. 2(b) we model $η_{2e}(E_C) = η_{0,C} e^{-a/E_C}$, motivated by the dependence on the control pulse energy of the former model. For the fit parameters we find $η_{0,C} = 0.107(7)$ and $a = 1.56(9) \cdot 10^2$ pJ. Due to non-ideal etalon adjustments, $η_{0,C}$ is slightly reduced compared to $η_{2e}$. The dependence on the detuning $Δ$ shown in Fig. 2(c) is modeled by $η_{2e}(Δ) = η_{0,Δ} e^{-α(Δ)}$, where $α(Δ)$ is a Lorentzian, reflecting that the memory efficiency is limited by collision-broadened absorption from the atomic ensemble. These models are not to be
FIG. 2. End-to-end efficiency ($\eta_{e2e}$) and signal-to-noise ratio (SNR). $\eta_{e2e}$ (blue circles) and SNR (green squares) of the memory as a function of signal pulse width $\Delta t_S$ (a), control pulse energy $E_C$ (b), and laser detuning $\Delta$ (c). The values selected for the measurements are $\Delta t_S = 25$ ns, $E_C = 560(50)$ pJ, and $\Delta = 2300(100)$ MHz red detuning from the atomic resonance, with only the respective parameter being varied.

FIG. 3. Noise analysis. (a) Noise detected on the APD, while scanning one of the etalons used for spectral filtering over two free spectral ranges (FSR) (shaded area corresponds to one FSR), while keeping the other etalon on resonance with the signal frequency. Large peaks at $-16.4$ GHz and $9.2$ GHz detuning correspond to leaked control-laser light, and small peaks close to the signal frequency (and at $25.6$ GHz detuning) originate from atomic noise (fluorescence, SRS and FWM). Vertical dashed lines indicate the signal (blue) and control (red) frequencies and vertical black solid lines indicate the limits of one FSR. (b) Exemplary measurement of total noise counts (the two different datasets were taken on different days) per retrieval attempt as a function of the energy in the control pulse $E_C$ for $\Delta = 334$ MHz (red detuning), and corresponding fits of the different noise components: fluorescence, spontaneous Raman scattering (SRS), and four-wave mixing (FWM). (c) Systematic study of the measured total noise counts and the corresponding fitted noise components at retrieval as a function of the control laser detuning $\Delta$. Negative values of $\Delta$ correspond to blue detuning from resonance. The noise is mainly dominated by SRS and fluorescence, while FWM is negligible in this setting.

taken as a result from a microscopic theory, but more as a hint to the underlying physical mechanisms. A full theoretical model is outlined in Refs. [42–47]. The models for the SNR are determined as the ratio between the efficiency- and the noise models, the latter of which are described in the next section.

To understand the limiting noise sources, a thorough experimental noise analysis is performed. When the temporal, spatial, and spectral leakage of the control light is sufficiently suppressed, the three most relevant noise sources in warm vapor quantum memories are fluorescence, SRS, and FWM [9]. Figure 3(a) shows the frequency dependence of the measured noise count rate, obtained by scanning the second etalon of the filtering system over two FSR, each of 25.6 GHz, by temperature tuning, while keeping the first etalon on resonance with the signal frequency. This measurement is taken with blocked signal laser. Two large peaks at $-16.4$ GHz and $9.2$ GHz detuning appear, corresponding to leaked control laser light. The transmitted control laser power is negligible when the etalon is in resonance with the signal, i.e. during storage experiments. Additionally, two smaller peaks close to the signal frequency appear approximately at 0 and 25.6 GHz detuning. This indicates that the only relevant noise sources during memory operation are of atomic origin. In order to
attain a better understanding of these, we measured
the number of noise photons \( N_{\text{noise}} \) counted on
the APD by varying \( \Delta \), while monitoring the control laser
frequency with the wavemeter. Here, \( \Delta = 0(60) \text{ MHz} \)
corresponds to the control transition. At each point, we
determined the noise count dependence on \( E_C \), yielding
a two dimensional analysis.

Figure 3(b) depicts the noise counts as a function
of the control pulse energy \( E_C \) at \( \Delta = 334 \text{ MHz} \)
(red detuning). The data can be modeled by a
second-order polynomial plus a saturating component,
\[
N_{\text{noise}} = b E_C^2 + c E_C + d E_C / (e + E_C),
\]
where the quadratic and linear contributions account for FWM
and SRS respectively, and the saturating part corre-
sponds to fluorescence noise. For the shown detuning
value we find \( b = 0(10^{-8}) \text{ (pJ)}^{-2}, c = 4(2) \cdot 10^{-5} \text{ (pJ)}^{-1},
\)
\( d = 7(2) \cdot 10^{-3} \text{ (pJ)}^{-1}, \) and \( e = 16(16) \text{ pJ}, \) indicating that
both fluorescence and SRS noise have a significant
contribution, while FWM vanishes.

Figure 3(c) depicts the behavior of the noise for varying
detuning \( \Delta \) at maximal \( E_C \). The error bars of the total
noise are of statistical origin. From fitting the different
noise components as a function of the control field
energy at each detuning value (as in Figure 3(b)), the
corresponding components fluorescence, SRS, and FWM
are determined at each point. Here, the error bars arise
from the fit error of the \( E_C \) dependence. The error bars at
\( \Delta = -526 \) and \( 774 \text{ MHz} \) are comparatively large,
since the fit hardly converged and FWM was manually excluded for these data points. From the analysis we
assume that FWM is negligible in the experiment, SRS
yields mainly a constant contribution to the noise, and
the detuning dependence of the total noise arises mainly
from fluorescence. The fluorescence was modeled as
a Voigt distribution with a Gaussian component of
380 MHz width at FWHM due to Doppler broadening,
and a Lorentzian component of 920 MHz at FWHM,
arising from power broadening by the control light and
pressure broadening due to collisions with the buffer
gas, with the only free parameter being the height. This
simple model conveys the general trend and is valid for
the here investigated small detunings on the order of
the broadened atomic transition. At large detunings, a
decay of SRS is expected, but even for the largest here
measured detuning values it is not pronounced enough
to cause significant deviations from the model. The
negligible FWM contribution is particularly interesting,
as this type of noise is commonly understood to be the
main drawback of vapor cell quantum memories and
efforts are made for its active mitigation [18, 22]. As we
operate with low bandwidth pulses, moderate control
powers, and at the D1 line, we manage to practically
suppress this type of noise. At detunings close to
zero, a local minimum appears in the noise data. We
presume this minimum to be caused by re-absorption
of the light due to residual spatial miss-match of the
beams along the whole length of the vapor cell. A
more detailed model considering this re-absorption,
the decay of SRS at large detunings, and the slight
fluctuations of the noise data has too many free param-
eters to fit the number of existing data points. More
investigation would be required to fully understand this
fact, and will be pursued in further experiments and
theoretical calculations. The resulting total noise model
\[
N_{\text{noise}}(\Delta) = N_{\text{SRS}} + N_{\text{FWM}} \cdot V(\Delta) + N_{\text{FWM}} \text{ is then a sum of the constant SRS contribution and the Voigt distribution resulting from fluorescence. For completeness, we also add the FWM contribution to the model. The resulting parameters are } N_{\text{SRS}} = 14(1) \cdot 10^{-3}, N_{\text{FWM}} = 7(2) \cdot 10^{-3}, \	ext{ and } N_{\text{FWM}} = 0(1) \cdot 10^{-3}. \text{ It is thus essential to reduce the former two for future optimization. Combining the models for } \eta_{\text{sel}} \text{ with the noise models yields the SNR models shown in Fig. 2(a)-(c). These are in reasonable agreement with the measured data, and confirm the applicability of the former models.}

In summary, we have realized, to the best of our
knowledge, the first single-photon-level storage in a
warm hyperfine EIT cesium memory on the Cs D1
line with an end-to-end efficiency of \( \eta_{\text{sel}} = (13 \pm 2)\% \),
which translates to an internal memory efficiency of
\( \eta_{\text{mem}} = (33 \pm 1)\% \). These values approach those needed for practical implementations on QRs [5, 6]. The simultaneously obtained SNR is \( 14 \pm 2 \), which can be converted to a noise level corresponding to \( \mu_1 = 0.07(2) \) signal photons. As the limiting noise sources, SRS
and fluorescence are identified. The present memory
is not yet fully optimized for long storage times, but
the geometry with parallel signal and control beams
in principle allows for storage times not limited by motional
dephasing, at least in the ms regime. Nonetheless, our
device can already find a use in quantum computation
and sensing applications where deterministic generation
of multiphoton states is required [48]. The used control
powers in this experiment were comparatively low. How-
ever, much higher powers on the order of hundreds of
mW will be needed in future experiments, where larger
beam sizes are planned, in order to further extend the
storage times. For minimizing SRS noise, the pumping
efficiency should be increased using higher pump laser
powers and larger beams. Further noise reduction is
planned by exploiting the Zeeman-substructure of the
cesium hyperfine states and preparing the atoms in the
\( m_F = +4 \) sublevel with \( \sigma^+ \) polarized light, thus
pressing residual FWM by selection rules, and reducing
possible SRS noise contributions [20, 49, 50]. In this
new configuration, storage times of a few milliseconds
are within reach. They will thus be extended to reach the
benchmark \( \tau_{\text{store}} = 1 \text{ ms} \) for practical applications in
QRs, by increasing the beam diameter up to the
cell diameter, hence inhibiting atomic diffusion out of
the optical interaction volume. By use of optimal
control pulses as in Ref. [21], we will simultaneously
boost the memory efficiency even further. With this, end-to-end efficiencies above 70% and simultaneously even higher signal-to-noise ratios, as well as ms-long storage times are within reach, laying the foundations for memory-assisted, satellite-suited QRs for long-distance quantum communication.

The authors kindly thank Gianni Buser for interesting and fruitful discussions. We acknowledge the support by the German Space Agency (DLR) with funds provided by the Federal Ministry of Economics and Technology (BMWi) under grant number 50RP2090 (QuMSeC) and by the Federal Ministry of Education and Research (BMBF) under grant number 16KISQ040K (Q-ToRX). M.G. acknowledges funding from the European Union’s Horizon 2020 research and innovation program under grant number 101039787 (Q-ToRX). E.R. acknowledges funding through the Helmholtz International Berlin Research School in Data Science (HEIBRiDS) and by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMWi) under grant number 50RP2090 (QuMSeC) provided by the Federal Ministry of Economics and Technology (BMWi) under grant number 16KISQ040K (Q-ToRX).
[38] D. F. Phillips, A. Fleischhauer, A. Mair, R. L. Walsworth, and M. D. Lukin, Phys. Rev. Lett. 86, 783 (2001).
[39] Y.-C. Tseng, Y.-C. Wei, and Y.-C. Chen, Opt. Express 30, 19944 (2022).
[40] D. A. Steck, Cesium D line data, https://steck.us/alkalidata/cesiumnumbers.pdf (Revision 2.2.1, 21 November 2019).
[41] Y.-F. Hsiao, P.-J. Tsai, H.-S. Chen, S.-X. Lin, C.-C. Hung, C.-H. Lee, Y.-H. Chen, Y.-F. Chen, I. A. Yu, and Y.-C. Chen, Phys. Rev. Lett. 120, 183602 (2018).
[42] A. V. Gorshkov, A. André, M. Fleischhauer, A. S. Sørensen, and M. D. Lukin, Phys. Rev. Lett. 98, 1 (2007).
[43] A. V. Gorshkov, A. André, M. D. Lukin, and A. S. Sørensen, Phys. Rev. A 76, 033804 (2007).
[44] A. V. Gorshkov, A. André, M. D. Lukin, and A. S. Sørensen, Phys. Rev. A 76, 033805 (2007).
[45] A. V. Gorshkov, A. André, M. D. Lukin, and A. S. Sørensen, Phys. Rev. A 76, 033806 (2007).
[46] A. V. Gorshkov, T. Calarco, M. D. Lukin, and A. S. Sørensen, Phys. Rev. A 77, 043806 (2008).
[47] M. T. Rakher, R. J. Warburton, and P. Treutlein, Phys. Rev. A 88, 053834 (2013).
[48] J. Nunn, N. K. Langford, W. S. Kolthammer, T. F. M. Champion, M. R. Sprague, P. S. Michelberger, X.-M. Jin, D. G. England, and I. A. Walmsley, Phys. Rev. Lett. 110, 133601 (2013).
[49] P.-C. Guan, Y.-F. Chen, and I. A. Yu, Phys. Rev. A 75, 013812 (2007).
[50] Z. Xu, Y. Wu, L. Tian, L. Chen, Z. Zhang, Z. Yan, S. Li, H. Wang, C. Xie, and K. Peng, Phys. Rev. Lett. 111, 240503 (2013).