Unified Treatment of Light Emission by Inelastic Tunneling: Interaction of Electrons and Photons beyond the Gap

Unai Muniain,[1] Ruben Esteban,[1,2] Javier Aizpurua,[1,3,4] and Jean-Jacques Greffet[5,*]

[1]*Donostia International Physics Center (DIPC), Paseo Manuel de Lardizabal 4, 20018 San Sebastián-Donostia, Basque Country, Spain*
[2]*Material Physics Center, CSIC-UPV/EHU, Paseo Manuel de Lardizabal 5, 20018 San Sebastián-Donostia, Basque Country, Spain*
[3]*IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Basque Country, Spain*
[4]*Department of Electricity and Electronics, University of the Basque Country, 20018 San Sebastián-Donostia, Basque Country, Spain*
[5]*Université Paris-Saclay, Institut d'Optique Graduate School, CNRS, Laboratoire Charles Fabry, 2 avenue A. Fresnel, 91127 Palaiseau, France*

(Received 1 October 2023; revised 24 January 2024; accepted 27 March 2024; published 24 April 2024)

A direct current through a metal-insulator-metal tunneling junction emits light when surface-plasmon polaritons (SPPs) are excited. Two distinct processes are believed to coexist in this light emission mediated by surface plasmons: inelastic tunneling, where electrons excite SPPs in the insulator gap, and hot-electron radiative decay, which occurs in the electrodes after elastic tunneling. Previous theoretical approaches to study light emission by inelastic tunneling have relied on Bardeen's approximation where the electronic wave functions are considered only in the barrier of the junction. In this work, we introduce an extension to models of inelastic tunneling by incorporating the full quantum device solution of the Schrödinger equation, which can also account for processes in the metallic electrodes. The extension unveils the existence of long-range correlations of the current density across the barrier and enables us to establish the equivalence between two models widely used in the past: (i) a calculation of the inelastic transition rate between two states across the barrier based on Fermi's golden rule and (ii) a calculation of the power transferred to plasmons by current fluctuations. Importantly, the new model accounts for processes that take place in the metallic electrodes and that could not be described within Bardeen's approximation. Hence, it is no longer necessary to invoke a hot-electron mechanism to obtain a dependence on the geometry of metallic electrodes. The new framework enables to discuss the role of surface plasmons localized in different metal-insulator interfaces and to include possible nonlocal effects at the interfaces.

DOI: 10.1103/PhysRevX.14.021017

Subject Areas: Condensed Matter Physics, Nanophysics, Plasmonics

I. INTRODUCTION

The study of metal-insulator-metal (MIM) junctions in vacuum as a source of electromagnetic radiation has grown significantly since the pioneering experiment carried out by Lambe and McCarthy almost half a century ago [1]. In these systems, light emission originates from the excitation of surface-plasmon polaritons (SPPs) by electronic injection when a bias potential is applied between the two planar metallic electrodes inducing a tunneling current in the insulator [see Fig. 1(a)]. The first studies of light emission by MIM junctions were carried out in planar junctions, which involved introducing some roughness on the surfaces of the electrodes to enable radiation of the SPPs. The role of the surface roughness to scatter the gap plasmons was further evidenced by depositing a disordered ensemble of silver nanoparticles on the upper metallic electrode to enhance the plasmon scattering [2]. Subsequently, the role of fast plasmons localized at the vacuum-metal interface, as opposed to the gap plasmon, was studied by several authors [3–6]. The interest in light emission by inelastic tunneling was revived when light emission from a scanning tunneling microscope (STM) was first measured [7] and the role of localized plasmons was shown [8]. This localized light emission process has been used as a spectroscopic tool to characterize samples with nanometer, and even subnanometric, spatial resolution by measuring the spatial variations of the emitted light [9–13]. After these milestone contributions, light emission in the tunneling regime has

*Corresponding author: jean-jacques.greffet@institutoptique.fr

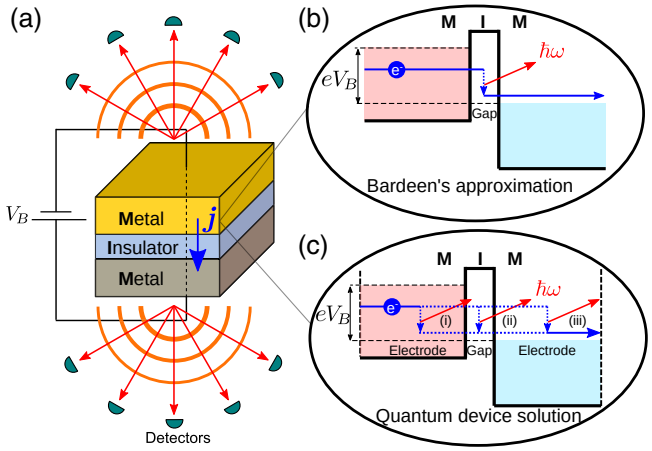

FIG. 1. Schematics of the light emission process from a MIM tunneling junction due to inelastic tunneling. (a) Sketch of a MIM junction. When a bias potential V_B is applied, an electronic current density $\mathbf{j}$ is induced in the system. An electron excites a plasmon of energy $\hbar\omega$ due to the inelastic tunneling mechanism. This plasmon decays radiatively by emitting light toward the detectors. (b) Sketch of the inelastic tunneling processes considered by Bardeen's approximation, where the electron decays only in the insulator gap. (c) Sketch of the inelastic tunneling processes considered by the quantum device solution (QDS) that we propose in this paper. The QDS includes inelastic tunneling processes that may occur (i) in the first metallic electrode, (ii) in the insulator gap, or (iii) in the second electrode.

been a topic of interest because of its many interesting features for potential applications. It can be used to design light sources with a very small footprint; it can be driven electrically and can be modulated at high frequency. Many recent experimental works have shown how to control the emission process, of either photons or plasmons, by controlling the environment using different types of resonant structures [14–21].

Despite the many potential applications of these systems as optical sources, their development has been hampered by the extremely low electron-to-SPP conversion rate. This rate is typically on the order of 10^{-5} for planar junctions [2]. It was found later that, with localized plasmons at the tip of a STM [22] or between the edges of two cubes [18], the conversion rates are, respectively, on the order of 10^{-3} and 10^{-2}. To explain light emission by tunneling junctions, it has been proposed that inelastic tunneling takes place in the insulator gap. This process is illustrated in Fig. 1(b). When a bias voltage V_B is applied, electrons (with charge $-e$) in one electrode gain an excess energy of $e|V_B|$. These electrons can tunnel through the few-nanometer-thick insulator gap between the metals to occupy an unoccupied state in the other electrode with lower energy. The excess energy excites a surface plasmon (with energy $\hbar\omega < e|V_B|$). The plasmon subsequently relaxes by either absorption or radiation. Since the dominant tunneling process is elastic tunneling generating a direct current [23,24], the efficiency of the inelastic process is expected to be low. It can be improved by reducing the elastic current [25,26].

It is worth mentioning that an effect similar to light emission by inelastic tunneling has been observed in the gigahertz regime when studying quantum electronic transport through very small tunneling junctions. This effect, called dynamical Coulomb blockade of tunneling, is a quantum effect in which tunneling of charge through a small junction is modified by the electromagnetic environment [27,28], which offers radiative decay channels. Hence, tunneling assisted by emission of gigahertz photons into the lines connected to the junction becomes possible. The direct observation of emission has been first reported in 2011 [29] in the 10 GHz range. The efficiency of this process can be unity when using superconductors.

Although models based on the inelastic tunneling mechanism describe many experimental data, they cannot account for all the reported experiments. For instance, Kirtley *et al.* measured light emission from metallic gratings. In this case, the theory based on inelastic tunneling underestimated the emitted power by an order of magnitude [4,19]. Furthermore, it was observed that the decay of the radiative intensity with the thickness of the top electrode could not be explained. It was also observed that emission can be quenched when introducing adsorbants on the air-metal interface of the electrode [30], an effect that cannot be explained if emission takes place in the barrier. To explain these discrepancies between theory and experiment, Kirtley *et al.* suggested an alternative mechanism of light emission, known as hot-electron decay [31]. In this mechanism, electrons first tunnel elastically and become hot electrons in the second electrode. Then, they thermalize through interactions with other electrons and with phonons. Because of the continuous pumping, a population of hot electrons, not described by a Fermi-Dirac distribution, is maintained. These hot electrons can relax by emitting surface plasmons at the air-metal interface. To support this alternative mechanism, Kirtley *et al.* performed photoluminescence measurements from metals under continuous laser pumping which produces hot electrons [32]. They observed similarities of the emitted light under electrical or optical excitation supporting the hot-electron mechanism. However, in contrast with the variety of models proposed to calculate the inelastic tunneling rate, the qualitative hot-electron emission process is still awaiting the implementation of a quantitative theoretical model that could be compared with experiments. To our knowledge, there is only a qualitative estimation of light emission in a STM including this hot-electron contribution, which found that the inelastic tunneling mechanism overcomes the hot-electron mechanism by a factor of approximately 10^3 [33]. In summary, the hot-electron picture has mostly remained a qualitative mechanism, and there is no available model able to explain the underestimation of the emitted power [4,19],

the emitted power dependence on the electrode thickness [4], and the emission quenching by adsorbants [30].

At this point, we emphasize that many authors have considered that emission originated in the metallic electrode is necessarily due to the hot-electron process [31,33] and that inelastic tunneling implies emission originated in the gap only [Fig. 1(b)] [31,33,34]. In what follows, we revisit this point of view by considering a more rigorous description of the inelastic tunneling that indicates how processes in the metal can also lead to emission, as depicted schematically in Fig. 1(c), without requiring the hot-electron mechanism. This new point of view enables to compute quantitatively the emitted power.

As stated, up to now, all the calculations (except Refs. [25,35]) assume that the light emission induced by inelastic tunneling takes place in the barrier. This process is typically modeled based on Bardeen's theory of electron tunneling (Bardeen's approximation) [36]. In this work, we abandon Bardeen's model and the idea that light is emitted in the barrier. We introduce an extension of the models of inelastic tunneling that is obtained by solving the Schrödinger equation in the complete MIM device and that we denote the full quantum device solution (QDS). We note that a similar approach [25] has been used to account for resonant tunneling in a double barrier. The QDS solution accounts for processes in the metallic electrodes [Fig. 1(c)]. We show that it enables to reproduce several features observed experimentally that could not be reproduced by Bardeen's theory. This includes a measured emitted power larger than what is predicted [31], an exponential decay of the emitted power on the thickness of the electrodes with a decay length much larger than the optical skin depth as predicted by Bardeen's theory [31]. We also show the importance of light emission at the air-metal interface due to nonlocal effects, a process that can be quenched by adsorbants on the air-metal interface of the electrode [30].

Additionally, within the QDS, we show the equivalence of the two main approaches that have been used to describe light emission induced by inelastic tunneling. The first one is based on the usual picture of radiative emission due to an electronic transition between two states. Fermi's golden rule can then be used to compute the rate of excitation of SPPs [37,38]. The second one is based on the classical picture of radiation due to time-dependent currents. It relies on the calculation of the power radiated by the time-dependent fluctuations of the current density [35,39]. Laks and Mills used this viewpoint to model the experiment of Lambe and McCarthy by calculating the emission efficiency of planar junctions with surface roughness in Ref. [35]. A summary of these two methods can be found in the review paper by Parzefall and Novotny [38]. To our knowledge, no systematic explicit proof of their equivalence has been reported. Here, the cross-spectral density of the current density is derived in a second quantization framework within the QDS. This new approach unveils an unexpected long-range correlation of the current density across the gap. Using this form of the cross-spectral density, it is possible to establish the equivalence of both methods.

The structure of this work is as follows. In Sec. II, we introduce the theoretical formulation of inelastic tunneling. We first (Sec. II A) present the QDS and use it to explain elastic tunneling phenomena, comparing the obtained results with those given by Bardeen's approximation. Then, in Sec. II B, we establish the equivalence of the two different approaches used to describe light emission by inelastic tunneling, based either (i) on the transition rate of electrons between two states or (ii) on the calculation of emission by fluctuating currents. In Sec. III, we calculate the radiative and nonradiative power due to SPP excitation in planar junctions. We show that the methods based on the QDS can explain available experimental observations. Finally, in Sec. IV, we summarize the conclusions and discuss further applications of the theory for future work.

II. DESCRIPTIONS OF LIGHT EMISSION FROM TUNNELING JUNCTIONS

We start by reviewing and comparing the frameworks that are used to describe electronic transport in a barrier. We first compare Bardeen's description of elastic tunneling in MIM junctions with a textbook solution of the Schrödinger equation denoted QDS. We then compute the inelastic tunneling rate using two models. We either use Fermi's golden rule or compute the power transferred from the fluctuating currents to the surface plasmons. To proceed, we derive and analyze the current density correlation function. We then establish the equivalence of the two models.

A. Elastic tunneling

In order to describe the dynamics of the electrons according to the theory of elastic electron tunneling, we define the electronic Hamiltonian $\hat{H}_{\mathrm{el}}$ as

$$\hat{H}_{\mathrm{el}} = \frac{-\hbar^2\nabla^2}{2m_{\mathrm{eff}}} + U(z), \tag{1}$$

where m_{eff} refers to the effective mass of the electrons in the MIM junction. This Hamiltonian is essentially a free electron model supplemented with the description of a potential to account for the barrier. It includes the kinetic energy of the electrons in the metals (first term in the right-hand side) and the potential energy $U(z)$ (second term in the right-hand side) [40].

To characterize the potential energy $U(z)$ [see sketch in the inset in Fig. 2(e)], we first consider that the metals placed on the left and on the right of the insulator gap have Fermi energies E_F^L and E_F^R, respectively. If there is no bias potential applied ($V_B = 0$), the system is at equilibrium and the Fermi surfaces of both metals are at the same energy,

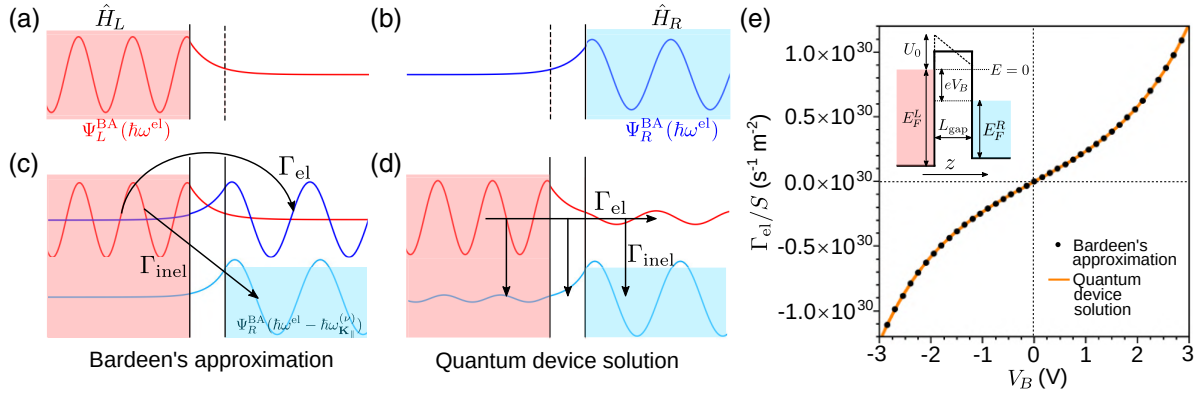

FIG. 2. Comparison of Bardeen's and QDS approaches for electron tunneling mechanisms. (a) Electronic wave function Ψ_L^{BA} corresponding to the left Hamiltonian $\hat{H}_L$ within Bardeen's approximation. (b) Electronic wave function Ψ_R^{BA} corresponding to the right Hamiltonian $\hat{H}_R$ within Bardeen's approximation. In (a) and (b), solid lines indicate the boundary between the metal considered in the Hamiltonian $\hat{H}_L$ or $\hat{H}_R$ and the gap, whereas dashed lines show the boundary between the gap and the metal that is considered absent in the corresponding Hamiltonian. (c),(d) Schematics of the processes of elastic tunneling (Γ_{el}) and inelastic tunneling (Γ_{inel}) using the wave functions corresponding to Bardeen's approximation (c) and the QDS (d). The wave functions are vertically shifted by their energy, where $\hbar\omega^{\mathrm{el}}$ is the initial electronic energy and $\hbar\omega_{\mathbf{K}_\parallel}^{(\nu)}$ is the energy of a SPP excited by the electron due to inelastic tunneling. (e) Elastic tunneling rate Γ_{el} per unit of surface area S of the metallic interfaces, calculated with Bardeen's approximation (black dots) and the QDS (orange line), plotted as a function of the bias potential V_B. The inset shows the schematics of the potential energy $U(z)$ of electrons in the MIM junction and the occupied states in each metal. In this inset, the dashed lines represent the potential energy $U(z)$ with the linear function $U_{\mathrm{gap}}(z)$ for the gap, while solid lines represent the potential energy $U(z)$ under the rectangular approximation that we consider in our calculations. The system considered in this figure is composed by an aluminum and a gold electrode separated by an Al_2O_3 insulator layer of thickness $L_{\mathrm{gap}} = 1$ nm.

which we set as the zero energy level. From this reference, the lowest values of the potential energy that electrons can have in the left and right metals are $-E_F^L$ and $-E_F^R$, respectively. The height of the barrier in the gap region is U_0. For applied bias potentials, the energies of all electrons in one metal are shifted by a value eV_B as compared to those in the other electrode. We maintain the zero energy reference in the Fermi energy of the left metal, and, thus, the minimum potential energy remains here as $U_L = -E_F^L$. In the right metal, this minimum energy shifts to the value $U_R = -E_F^R - eV_B$, while the energy of the highest occupied state becomes $-eV_B$. This shift of the energy levels also affects the work functions of the metals, which causes a modification of the potential inside the insulator barrier. The potential in this region can be approximated with the linear function $U_{\mathrm{gap}}(z) = U_0 - eV_B(z/L_{\mathrm{gap}})$, where z is the perpendicular direction to the interfaces of the junction and L_{gap} is the thickness of the insulator gap [the potential $U(z)$ with the linear function is indicated by dashed lines in the inset in Fig. 2(e)]. However, to obtain simple analytical expressions of the electronic wave functions, we use a rectangular approximation for the barrier potential, which takes a constant value determined by the average $U_{\mathrm{gap}}(z) \approx U_0 - (eV_B/2)$ [we show the rectangular potential $U(z)$ in the inset in Fig. 2(e) by solid lines].

With this model of the MIM junction, the electronic properties of the system are obtained by solving the Schrödinger equation of Eq. (1). For a fixed energy $\hbar\omega^{\mathrm{el}}$ and parallel wave vector $\mathbf{k}_\parallel$, the Hamiltonian of a rectangular barrier contains two degenerate states, of the form

$$\Psi_L^{\mathrm{QDS}}(\mathbf{r}) = \begin{cases} \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}}\left(e^{ik_{zL}z} + r_L e^{-ik_{zL}z}\right)e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & z \le 0, \\ \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}}\left(\alpha_L e^{k_{z\mathrm{gap}}z} + \beta_L e^{-k_{z\mathrm{gap}}z}\right)e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & 0 < z \le L_{\mathrm{gap}}, \\ \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}}\, t_L e^{ik_{zR}z}e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & L_{\mathrm{gap}} < z \end{cases} \tag{2}$$

and

$$\Psi_R^{\mathrm{QDS}}(\mathbf{r}) = \begin{cases} \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}} t_R e^{-ik_{zL}(z-L_{\mathrm{gap}})} e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & z \le 0, \\ \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}} (\alpha_R e^{-k_{z\mathrm{gap}}(z-L_{\mathrm{gap}})} + \beta_R e^{k_{z\mathrm{gap}}(z-L_{\mathrm{gap}})}) e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & 0 < z \le L_{\mathrm{gap}}, \\ \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}} (e^{-ik_{zR}(z-L_{\mathrm{gap}})} + r_R e^{ik_{zR}(z-L_{\mathrm{gap}})}) e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & L_{\mathrm{gap}} < z. \end{cases} \tag{3}$$

L_z is the (arbitrary) length of the system in the z direction, whereas S is the surface of the interfaces in the $\mathbf{r}_\parallel = (x, y)$ plane. These two parameters act as normalization constants in the wave functions. Furthermore,

$$k_{zL(R)}(\omega^{\mathrm{el}}, \mathbf{k}_\parallel) = \sqrt{\frac{2m_{\mathrm{eff}}}{\hbar^2}(\hbar\omega^{\mathrm{el}} - U_{L(R)}) - |\mathbf{k}_\parallel|^2} \tag{4}$$

is the z component of the wave vector of an electron in the left (right) metal. The spatial decay of the wave function in the gap region is governed by the value

$$k_{z\mathrm{gap}}(\omega^{\mathrm{el}}, \mathbf{k}_\parallel) = \sqrt{\frac{2m_{\mathrm{eff}}}{\hbar^2}(U_{\mathrm{gap}} - \hbar\omega^{\mathrm{el}}) + |\mathbf{k}_\parallel|^2}. \tag{5}$$

The coefficients $\alpha_{L(R)}$, $\beta_{L(R)}$, $r_{L(R)}$, and $t_{L(R)}$ are given in Appendix A. The wave functions $\Psi_L^{\mathrm{QDS}}(\mathbf{r})$ in Eq. (2) and $\Psi_R^{\mathrm{QDS}}(\mathbf{r})$ in Eq. (3) are the textbook solutions to describe quantum tunneling from the left metal to the right metal and vice versa, respectively. $r_{L(R)}$ and $t_{L(R)}$ give the reflected and transmitted amplitudes, respectively. Since the wave functions are calculated in the complete MIM device, we refer to these wave functions as the QDS.

These expressions can be simplified in an approximation first carried out by Bardeen [36] that is widely used in the study of elastic tunneling in MIM junctions. Under Bardeen's approximation (BA), the two metals are considered as two separate entities, and the electronic states of each metal are not affected by the other one. We illustrate in Figs. 2(a) and 2(b) with vertical solid lines the metal-insulator boundary considered in the Hamiltonians $\hat{H}_L$ and $\hat{H}_R$, respectively. The eigenstates of the system are obtained by solving the Schrödinger equation with the Hamiltonians $\hat{H}_L$ and $\hat{H}_R$ separately. Each of these Hamiltonians includes operators for the kinetic and the potential energy. The potential energy operator considers the energy level of the corresponding metal (U_L or U_R) and of the barrier (U_{gap}). However, when calculating the wave function associated with each electrode, we do not consider the presence of the other one; i.e., we extend the gap barrier to infinity [we show in Figs. 2(a) and 2(b) the metal-insulator boundary neglected in each Hamiltonian by dashed lines]. Following this procedure, we obtain the wave functions of the left metal [see Fig. 2(a)] [41]:

$$\Psi_L^{\mathrm{BA}}(\mathbf{r}) = \begin{cases} \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}} \left(e^{ik_{zL}z} + \frac{ik_{zL}+k_{z\mathrm{gap}}}{ik_{zL}-k_{z\mathrm{gap}}} e^{-ik_{zL}z} \right) e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & z \le 0, \\ \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}} \frac{2ik_{zL}}{ik_{zL}-k_{z\mathrm{gap}}} e^{-k_{z\mathrm{gap}}z} e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & z > 0. \end{cases} \tag{6}$$

Equivalently, the wave functions corresponding to the right metal have the form [see Fig. 2(b)]

$$\Psi_R^{\mathrm{BA}}(\mathbf{r}) = \begin{cases} \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}} \frac{2ik_{zR}}{ik_{zR}-k_{z\mathrm{gap}}} e^{k_{z\mathrm{gap}}(z-L_{\mathrm{gap}})} e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & z \le L_{\mathrm{gap}}, \\ \frac{1}{\sqrt{L_z}}\frac{1}{\sqrt{S}} \left(e^{-ik_{zR}(z-L_{\mathrm{gap}})} + \frac{ik_{zR}+k_{z\mathrm{gap}}}{ik_{zR}-k_{z\mathrm{gap}}} e^{ik_{zR}(z-L_{\mathrm{gap}})} \right) e^{i\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel} & z > L_{\mathrm{gap}}. \end{cases} \tag{7}$$

Under Bardeen's approximation, the interaction between the two metals is treated at a perturbative level. Each electron of the left metal is at first in the state $|\Psi_L^{\mathrm{BA}}\rangle$ corresponding to the Hamiltonian $\hat{H}_L$. The rest of the electronic Hamiltonian of Eq. (1), $\hat{H}_{\mathrm{el}} - \hat{H}_L$, induces transitions to states of the form $|\Psi_R^{\mathrm{BA}}\rangle$, as schematically shown in Fig. 2(c) by the arrow labeled with the symbol Γ_{el}. The assumption of weak tunneling due to a sufficiently thick gap implies that the transition rate between two particular left and right states is given by Fermi's golden rule, as (see derivation in Appendix A)

$$\begin{aligned} \Gamma_{L\to R}^{\mathrm{BA}} &= \frac{2\pi}{\hbar^2}\delta(\omega_L^{\mathrm{el}} - \omega_R^{\mathrm{el}}) |\langle \Psi_R^{\mathrm{BA}} | \hat{H}_{\mathrm{el}} - \hat{H}_L | \Psi_L^{\mathrm{BA}} \rangle|^2 \\ &= \frac{(2\pi)^3\hbar^2}{4m_{\mathrm{eff}}^2} S\delta(\omega_L^{\mathrm{el}} - \omega_R^{\mathrm{el}})\delta(\mathbf{k}_{\parallel R} - \mathbf{k}_{\parallel L}) \left| \Psi_R^{\mathrm{BA}*}(z) \frac{\partial \Psi_L^{\mathrm{BA}}(z)}{\partial z} - \Psi_L^{\mathrm{BA}}(z) \frac{\partial \Psi_R^{\mathrm{BA}*}(z)}{\partial z} \right|_{z=L_{\mathrm{gap}}}^2. \end{aligned} \tag{8}$$

By considering a single incident electron from the left metal, its tunneling rate Γ_L^{BA} is given by the sum over all possible unoccupied states $|\Psi_R^{\mathrm{BA}}\rangle$ of the right metal, i.e., $\Gamma_L^{\mathrm{BA}} = \sum_{\mathbf{k}_R} \Gamma_{L\to R}^{\mathrm{BA}}$. To characterize these electronic states in the metals, we use the periodic boundary conditions so that the states are given by wave vectors of the form $\mathbf{k}_{L(R)} = [(2\pi/\sqrt{S})n_x, (2\pi/\sqrt{S})n_y, (2\pi/L_z)n_z]$ with integers n_x, n_y, and n_z. In this context, we can substitute the discrete sums $\sum_{\mathbf{k}_{L(R)}}$ with the integrals $[L_z S/(2\pi)^3]\int d\mathbf{k}_{L(R)}$, and by performing the integral of $\Gamma_{L\to R}^{\mathrm{BA}}$ [with the wave functions of Eqs. (6) and (7)] over the right states indicated by $\mathbf{k}_R$, we obtain the tunneling rate per incident electron given by

$$\Gamma_L^{\mathrm{BA}} = \frac{\hbar}{m_{\mathrm{eff}} L_z} \frac{16 k_{zL}^2 k_{zR} k_{z\mathrm{gap}}^2}{(k_{zL}^2 + k_{z\mathrm{gap}}^2)(k_{zR}^2 + k_{z\mathrm{gap}}^2)} e^{-2k_{z\mathrm{gap}} L_{\mathrm{gap}}}. \tag{9}$$

The transition rate of Eq. (9) has been derived using Bardeen's approximation. We now turn to a more rigorous approach that does not use Bardeen's approximation. It is possible to obtain the tunneling rate using the QDS. Since the electrons under this description are already in an eigenstate of $\hat{H}_{\mathrm{el}}$ from the beginning, the tunneling properties are included in $\Psi_L^{\mathrm{QDS}}(\mathbf{r})$, as we indicate in Fig. 2(d) by the arrow with the Γ_{el} label. From this wave function, we can obtain its associated probability current density from the general definition [42]

$$j_z(\mathbf{r}) = \frac{i\hbar e}{2m_{\mathrm{eff}}}\left[\Psi^*(\mathbf{r})\frac{\partial \Psi(\mathbf{r})}{\partial z} - \Psi(\mathbf{r})\frac{\partial \Psi^*(\mathbf{r})}{\partial z}\right], \tag{10}$$

which for Eq. (2) is constant over space with the value $j_z(z) = -(1/L_z S)(\hbar e k_{zR}/m_{\mathrm{eff}})|t_L|^2$. Since this expression gives the amount of charge that crosses the junction per unit of area and time, the tunneling rate Γ_L is directly calculated as

$$\Gamma_L^{\mathrm{QDS}} = \frac{S}{-e} j_z. \tag{11}$$

By introducing the value of the coefficient t_L (given in Appendix A) into Eq. (2) and applying Eq. (10) to calculate the current density $j_z(z)$, we obtain

$$\Gamma_L^{\mathrm{QDS}} = \frac{\frac{\hbar}{m_{\mathrm{eff}} L_z} 16 k_{zL}^2 k_{zR} k_{z\mathrm{gap}}^2 e^{-2k_{z\mathrm{gap}} L_{\mathrm{gap}}}}{(1 + e^{-4k_{z\mathrm{gap}} L_{\mathrm{gap}}})(k_{z\mathrm{gap}}^2 + k_{zL}^2)(k_{z\mathrm{gap}}^2 + k_{zR}^2) - 2e^{-2k_{z\mathrm{gap}} L_{\mathrm{gap}}}[(k_{z\mathrm{gap}}^2 - k_{zL}^2)(k_{z\mathrm{gap}}^2 - k_{zR}^2) - 4k_{z\mathrm{gap}}^2 k_{zL} k_{zR}]}. \tag{12}$$

The elastic tunneling rate given by Bardeen's approximation [Eq. (9)] is identical to the expression that is obtained from the rate of the QDS [Eq. (12)] under the assumption of weak tunneling, i.e., for $k_{z\mathrm{gap}} L_{\mathrm{gap}} \gg 1$. To analyze the validity of this assumption, we focus on an Al-Al_2O_3-Au junction as a particular example, which is typically used to analyze elastic and inelastic tunneling phenomena. We use numerical values of the Fermi energies $E_F^L = 11.5$ eV and $E_F^R = 5.5$ eV for Al and Au, respectively [43]. Furthermore, it has been measured that the effective mass of the electrons in alumina junctions is $m_{\mathrm{eff}} = 0.23 m_e$ (where m_e is the electron mass) [44], and we fix the height of the barrier on a typical value of $U_0 = 2$ eV. Since Bardeen's approximation works accurately for $k_{z\mathrm{gap}} L_{\mathrm{gap}} \gg 1$, the largest mismatch with the QDS occurs in the regime of very thin layers and of high bias potentials, where $k_{z\mathrm{gap}}$ is smallest according to Eq. (5). However, even for values of $L_{\mathrm{gap}} = 1$ nm and $V_B = 3$ V, we have checked that Bardeen's approximation underestimates the tunneling rate Γ_L of the electrons at the Fermi surface only by a factor of 0.4%. Thus, Bardeen's approximation is well justified when considering elastic tunneling.

Although the tunneling rate per electron allows us to compare Bardeen's approximation with the QDS by means of an analytical expression [Eqs. (9) and (12)], the measurable quantity in experiments is the intensity of the electronic current (related to the total tunneling rate Γ_{el}) instead of the tunneling rate per incident electron Γ_L. In order to model these experiments, we calculate Γ_{el} by summing Eqs. (9) and (12) over all occupied initial states of the left metal that can tunnel to unoccupied states of the right metal $\Gamma_{\mathrm{el}} = \sum_{\mathbf{k}_L} \Gamma_L f_{\mathrm{FD}}^L(\mathbf{k}_L)[1 - f_{\mathrm{FD}}^R(\mathbf{k}_L)]$. To perform this calculation, we have included the probability that a state is occupied in the left metal and unoccupied in the right metal. Since the states described by Bardeen's approximation are localized in a single metal, these probabilities are dictated by the Fermi-Dirac occupation factors of its respective metal

$$f_{\mathrm{FD}}^{L(R)}(\mathbf{k}_L) = \left[1 + \exp\left(\frac{\hbar\omega^{\mathrm{el}}(\mathbf{k}_L) - E_F^{L(R)}}{k_B T}\right)\right]^{-1}$$

at temperature T, with Boltzmann constant k_B. The assignment of a Fermi-Dirac occupation factor is not so straightforward in the approach of the QDS, because the corresponding electronic states are delocalized over the two metals with different Fermi levels. However, following a similar argument than for Bardeen's approximation, we associate the occupation factor $f_{\mathrm{FD}}^L(\mathbf{k}_L)$ to the states $\Psi_L^{\mathrm{QDS}}(\mathbf{r})$ that originate from the left metal, and accordingly

the factor $[1-f_{\mathrm{FD}}^{R}(\mathbf{k}_L)]$ corresponds to the unoccupied states $\Psi_R^{\mathrm{QDS}}(\mathbf{r})$ that tunnel from the right to the left metal.

We plot in Fig. 2(e) the elastic tunneling rate Γ_{el} per surface area S as a function of the applied voltage V_B for temperature $T=0$ (we have checked that the following discussion remains also valid for room and larger temperatures). We compare the results obtained with Bardeen's approximation (black dots) and the QDS (orange line) for a thin gap of width $L_{\mathrm{gap}}=1$ nm. The two approaches follow a nearly identical trend with a relative error of at most 2.1×10^{-3} at $V_B=3$ V. In most experiments, the gap thickness L_{gap} is larger, which decreases the error between the two approaches even more. Therefore, the agreement between the Bardeen's and QDS approaches indicates that associating single-metal Fermi-Dirac occupation factors $f_{\mathrm{FD}}^{L(R)}(\mathbf{k}_L)$ to the states of the QDS works accurately, and, thus, we follow this methodology in the following subsection, where we turn to the inelastic tunneling current.

B. Inelastic tunneling

The phenomenon of light emission due to inelastic tunneling has been computed using two different models in the literature: a calculation based on Fermi's golden rule and a calculation of the radiated power due to current density fluctuations. These two models are based on the two usual physical pictures of light emission: The quantum matter picture is based on the radiative relaxation of an excited state, and the classical electromagnetic picture is based on the power radiated by a time-dependent current.

In this subsection, we aim to provide a comprehensive review of the two models and to prove their equivalence under appropriate conditions. Furthermore, these methods are often used within the framework of Bardeen's approximation, but we emphasize here the consequences of using them within the QDS. Finally, we compute the inelastic transition rate leading to surface-plasmon excitation, as this plasmonic contribution dominates the local density of electromagnetic states. Once this transition rate is known, multiplying it by the surface-plasmon radiative decay yield gives the photon emission rate.

1. Fermi's golden rule

In the description of light emission from tunneling junctions, it is necessary to include the interaction between electrons and electromagnetic modes. Among different points of view to account for this interaction, one of them is a direct extension of models of elastic tunneling that involves introducing the light-matter coupling in the quantum Hamiltonian of Eq. (1). The effects of this coupling are usually treated under the formalism of Fermi's golden rule. This general approach has been used in a large variety of systems, such as in the analysis of photon emission from superconducting junctions [28]. Focusing now on the specific case of planar MIM junctions, the complete Hamiltonian that describes elastic tunneling together with the interaction between electrons and SPPs is

$$\hat{H}_{\text{el-SPP}}=\hat{H}_{\mathrm{el}}+\sum_{\mathbf{K}_\parallel}\sum_{\nu}\hbar\omega_{\mathbf{K}_\parallel}^{(\nu)}\hat{a}_{\mathbf{K}_\parallel}^{\dagger(\nu)}\hat{a}_{\mathbf{K}_\parallel}^{(\nu)}-\frac{e}{m_{\mathrm{eff}}}\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}. \tag{13}$$

Together with the electronic Hamiltonian $\hat{H}_{\mathrm{el}}$ in Eq. (1), the second term on the right-hand side of this expression is the plasmonic Hamiltonian. The superscript ν refers to all different SPP modes of the system (mostly localized at different interfaces) whose dispersion $\omega_{\mathbf{K}_\parallel}^{(\nu)}$ is a function of the parallel component of the wave vector $\mathbf{K}_\parallel$. We include the corresponding creation operator $\hat{a}_{\mathbf{K}_\parallel}^{\dagger(\nu)}$ and annihilation operator $\hat{a}_{\mathbf{K}_\parallel}^{(\nu)}$ for each vector $\mathbf{K}_\parallel$. The last term corresponds to the light-matter interaction $\hat{H}_{\mathrm{int}}$, which, depending on the system and its mode structure, is described with the vector potential $\hat{\mathbf{A}}$ in the Coulomb gauge as $\hat{H}_{\mathrm{int}}=-(e/m_{\mathrm{eff}})\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}$ or in terms of the scalar potential $\hat{\phi}$ as $\hat{H}_{\mathrm{int}}=-e\hat{\phi}$ [33,37,45]. In this work, we use the former interaction term, because all transverse modes in planar junctions can be described entirely with the vector potential. The operator $\hat{\mathbf{p}}=-i\hbar\nabla$ acts on the electronic wave functions, whereas the field operator $\hat{\mathbf{A}}$ is written after the decomposition into all plasmonic modes as [46]

$$\hat{\mathbf{A}}(\mathbf{r},t)=\sum_{\mathbf{K}_\parallel}\sum_{\nu}\sqrt{\frac{\hbar}{2\varepsilon_0 S\omega_{\mathbf{K}_\parallel}^{(\nu)}}}e^{i\mathbf{K}_\parallel\cdot\mathbf{r}_\parallel}\mathbf{u}_{\mathbf{K}_\parallel}^{(\nu)}(z)\hat{a}_{\mathbf{K}_\parallel}^{(\nu)}e^{-i\omega_{\mathbf{K}_\parallel}^{(\nu)}t}+\sqrt{\frac{\hbar}{2\varepsilon_0 S\omega_{\mathbf{K}_\parallel}^{(\nu)}}}e^{-i\mathbf{K}_\parallel\cdot\mathbf{r}_\parallel}\mathbf{u}_{\mathbf{K}_\parallel}^{*(\nu)}(z)\hat{a}_{\mathbf{K}_\parallel}^{\dagger(\nu)}e^{i\omega_{\mathbf{K}_\parallel}^{(\nu)}t}, \tag{14}$$

where ε_0 is the vacuum permittivity and $\mathbf{u}_{\mathbf{K}_\parallel}^{(\nu)}(z)$ gives the spatial dependence of the vector potential along the z direction for each plasmonic mode, under the condition $i\mathbf{K}_\parallel\cdot\mathbf{u}_{\mathbf{K}_\parallel}^{(\nu)}(z)+(\partial/\partial z)(\mathbf{u}_{\mathbf{K}_\parallel}^{(\nu)}(z)\cdot\mathbf{n}_z)=0$ implied by the Coulomb gauge ($\mathbf{n}_z$ is the unit vector of the z direction). Furthermore, the quantization of each plasmonic mode with energy $\hbar\omega_{\mathbf{K}_\parallel}^{(\nu)}$ leads to the following normalization condition of $\mathbf{u}_{\mathbf{K}_\parallel}^{(\nu)}(z)$ [46,47]:

$$\int\frac{1}{2}\left\{\frac{\partial}{\partial\omega}[\omega\varepsilon(z,\omega)]|\mathbf{u}_{\mathbf{K}_\parallel}^{(\nu)}(z)|^2+\left|\frac{\varepsilon(z,\omega)}{|\mathbf{K}_\parallel|}\frac{\omega}{c}(\mathbf{u}_{\mathbf{K}_\parallel}^{(\nu)}(z)\cdot\mathbf{n}_z)\right|^2\right\}dz=1, \tag{15}$$

with c the speed of light in vacuum and where $\varepsilon(z,\omega)$ gives the spatial distribution of the permittivity along the z direction at frequency ω.

To quantize the plasmon field, losses have been neglected. This approximation is valid inasmuch as the

density of states is not perturbed significantly [46]. It provides an accurate description of the electron-plasmon coupling which will be used to compute the plasmon emission rate. The photon emission rate can be computed subsequently using the radiative yield of the plasmon. The Hamiltonian of Eq. (13) is not exactly solvable, in general, and the usual approach to describe inelastic tunneling is to treat $\hat{H}_{\rm int} = -(e/m_{\rm eff})\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}$ as a perturbative term under the assumption of weak light-matter interaction. It is first assumed that the electrons come from the left metal and that there is no excited plasmon. Therefore, the initial state is of the form $|\Psi_L\rangle \otimes |0_{\rm pl}\rangle$ (note that, if we do not specify the superscript BA or QDS in the electronic state $\Psi_{L(R)}$, the expression is valid for both of these approaches). Here, the state $|0_{\rm pl}\rangle$ implies that all plasmonic modes ν at all wave vectors $\mathbf{K}_\parallel$ are in the zero occupation number state. The final plasmonic state is of the form $|1^{(\nu)}_{\mathbf{K}_\parallel}\rangle$, where all modes are unoccupied except for a SPP mode ν of parallel wave vector $\mathbf{K}_\parallel$ with occupation number 1. On the other hand, the electronic part of the final state can be of the form $|\Psi_L\rangle$ or $|\Psi_R\rangle$, depending on the Fermi-Dirac occupation factor of these states. Because of the applied bias potential, it is expected that, at low temperatures, the number of unoccupied states will be significantly greater in the right metal than in the left metal for final energies lower than that of the initial state. Thus, the transitions to the states of the left metal are highly suppressed, i.e., $\Gamma_{L\to L} \ll \Gamma_{L\to R}$, and in this formalism we focus on only the transitions of the form $L \to R$. In other words, we consider only the recombination of an electron coming from the left to a hole coming from the right. Specifically, the tunneling rate according to Fermi's golden rule reads

$$\Gamma_{L\to R} = \frac{2\pi}{\hbar^2}\sum_{\mathbf{K}_\parallel}\sum_{\nu}\delta(\omega_L^{\rm el} - \omega_R^{\rm el} - \omega^{(\nu)}_{\mathbf{K}_\parallel})\left|\mathcal{M}^{(\nu)}_{L,R,\mathbf{K}_\parallel}\right|^2, \tag{16}$$

with the matrix element

$$\begin{aligned}\mathcal{M}^{(\nu)}_{L,R,\mathbf{K}_\parallel} &= \left\langle \Psi_R, 1^{(\nu)}_{\mathbf{K}_\parallel}\middle|\hat{H}_{\rm int}\middle|\Psi_L, 0_{\rm pl}\right\rangle = \sqrt{\frac{\hbar}{2\varepsilon_0 S\omega^{(\nu)}_{\mathbf{K}_\parallel}}}\frac{i\hbar e}{m_{\rm eff}}\int_{V_{\rm gap}+V_{\rm met}}\Psi_R^*(\mathbf{r})\left[\mathbf{u}^{(\nu)}_{\mathbf{K}_\parallel}(z)e^{-i\mathbf{K}_\parallel\cdot\mathbf{r}_\parallel}\right]\cdot\nabla\Psi_L(\mathbf{r})d\mathbf{r}\\ &= \left\langle \Psi_L, 0_{\rm pl}\middle|\hat{H}_{\rm int}\middle|\Psi_R, 1^{(\nu)}_{\mathbf{K}_\parallel}\right\rangle^* = \sqrt{\frac{\hbar}{2\varepsilon_0 S\omega^{(\nu)}_{\mathbf{K}_\parallel}}}\frac{-i\hbar e}{m_{\rm eff}}\int_{V_{\rm gap}+V_{\rm met}}\Psi_L(\mathbf{r})\left[\mathbf{u}^{(\nu)}_{\mathbf{K}_\parallel}(z)e^{-i\mathbf{K}_\parallel\cdot\mathbf{r}_\parallel}\right]\cdot\nabla\Psi_R^*(\mathbf{r})d\mathbf{r},\end{aligned} \tag{17}$$

where we specify that in the QDS this integral has to be performed in the volume of the gap $V_{\rm gap}$ and of the metallic regions $V_{\rm met}$. We stress that this integral accounts for the exact spatial dependence of both the electronic wave functions and the surface-plasmon modes in the gap and in the metal. Note that, when accounting for metallic layers with finite thickness, we limit the integration over z but, for the sake of simplicity, we still use the electron wave functions given by Eqs. (2) and (3) established for a semi-infinite reservoir. Since the contributions of the gap and the metals must be summed, these two terms can produce interferences, as we discuss in Sec. III. Furthermore, combining the two equivalent forms of the matrix element shown in Eq. (17), we can write $\mathcal{M}^{(\nu)}_{L,R,\mathbf{K}_\parallel}$ in a symmetric form that is similar to the elastic tunneling rate obtained with Bardeen's approximation Eq. (9) as

$$\mathcal{M}^{(\nu)}_{L,R,\mathbf{K}_\parallel} = \sqrt{\frac{\hbar}{2\varepsilon_0 S\omega^{(\nu)}_{\mathbf{K}_\parallel}}}\frac{i\hbar e}{2m_{\rm eff}}\int_{V_{\rm gap}+V_{\rm met}}\left[\mathbf{u}^{(\nu)}_{\mathbf{K}_\parallel}(z)e^{-i\mathbf{K}_\parallel\cdot\mathbf{r}_\parallel}\right]\cdot[\Psi_R^*(\mathbf{r})\nabla\Psi_L(\mathbf{r}) - \Psi_L(\mathbf{r})\nabla\Psi_R^*(\mathbf{r})]d\mathbf{r}, \tag{18}$$

where, in analogy with Eq. (10), the wave functions $\Psi_L(\mathbf{r})$ and $\Psi_R(\mathbf{r})$ appear in the form of the inelastic current density as [37]

$$\mathbf{j}_{L\to R}(\mathbf{r}) = \frac{i\hbar e}{2m_{\rm eff}}[\Psi_R^*(\mathbf{r})\nabla\Psi_L(\mathbf{r}) - \Psi_L(\mathbf{r})\nabla\Psi_R^*(\mathbf{r})]. \tag{19}$$

The experimentally measurable quantity is the intensity of the emitted light due to the total inelastic tunneling rate $\Gamma_{\rm inel}$, which is given by the sum over all occupied initial states in the left metal and unoccupied final states of the right metal, calculated as $\Gamma_{\rm inel} = \sum_{\mathbf{k}_L}\sum_{\mathbf{k}_R}\Gamma_{L\to R}f^L_{\rm FD}(\mathbf{k}_L)[1 - f^R_{\rm FD}(\mathbf{k}_R)]$:

$$\Gamma_{\rm inel} = \sum_{\mathbf{K}_\parallel}\sum_{\nu}\sum_{\mathbf{k}_L}\sum_{\mathbf{k}_R}\frac{2\pi}{\hbar^2}\delta\left(\omega_L^{\rm el} - \omega_R^{\rm el} - \omega^{(\nu)}_{\mathbf{K}_\parallel}\right)f^L_{\rm FD}(\mathbf{k}_L)[1 - f^R_{\rm FD}(\mathbf{k}_R)]|\mathcal{M}^{(\nu)}_{L,R,\mathbf{K}_\parallel}|^2. \tag{20}$$

The corresponding total power transferred by the tunneling current to the plasmons is given by

$$\mathcal{P} = \sum_{\mathbf{K}_\parallel}\sum_{\nu}\sum_{\mathbf{k}_L}\sum_{\mathbf{k}_R} \hbar\omega^{(\nu)}_{\mathbf{K}_\parallel}\Gamma_{L\to R}\left(\omega^{(\nu)}_{\mathbf{K}_\parallel}\right) f^L_{\mathrm{FD}}(\mathbf{k}_L)[1-f^R_{\mathrm{FD}}(\mathbf{k}_R)]$$
$$= \sum_{\mathbf{K}_\parallel}\sum_{\nu}\sum_{\mathbf{k}_L}\sum_{\mathbf{k}_R} \frac{2\pi}{\hbar^2}\delta\left(\omega^{\mathrm{el}}_L-\omega^{\mathrm{el}}_R-\omega^{(\nu)}_{\mathbf{K}_\parallel}\right) f^L_{\mathrm{FD}}(\mathbf{k}_L)[1-f^R_{\mathrm{FD}}(\mathbf{k}_R)]|\mathcal{M}_{L,R,\mathbf{K}_\parallel}|^2\hbar\omega^{(\nu)}_{\mathbf{K}_\parallel}. \quad (21)$$

Following the standard procedure of elastic tunneling, the calculation of these transition rates is usually done under Bardeen's approximation, by evaluating the integral of the matrix element Eq. (18) just in the gap with the wave functions from Eqs. (6) and (7). This approximation assumes that the SPP is excited in the gap due to the inelastic tunneling processes across the barrier. This approach also gives an intuitive understanding on how the optical properties of the MIM junction influence the light emission process. We first notice that, in typical gaps of a few nanometers, the variation of the electromagnetic field is very smooth and it can be considered as constant in the integration region of Eq. (18). This assumption implies that the electronic and optical properties of the junction can be considered separately in Eq. (16). On the one hand, $\Gamma_{L\to R}$ is proportional to the electronic matrix element $|\langle\Psi^{\mathrm{BA}}_R|\mathbf{p}|\Psi^{\mathrm{BA}}_L\rangle + \langle\Psi^{\mathrm{BA}}_L|\mathbf{p}|\Psi^{\mathrm{BA}}_R\rangle|^2 = |-i\hbar\int\Psi^{\mathrm{BA}*}_R(\mathbf{r})\nabla\Psi^{\mathrm{BA}}_L(\mathbf{r}) - \Psi^{\mathrm{BA}}_L(\mathbf{r})\nabla\Psi^{\mathrm{BA}*}_R(\mathbf{r})d\mathbf{r}|^2$. On the other hand, by doing the sum over all plasmonic modes, the tunneling rate is also proportional to the projected local optical density of states $\rho_{\mathrm{opt}}(\mathbf{r}) = \sum_{\mathbf{K}_\parallel}\sum_{\nu}\delta(\omega^{\mathrm{el}}_L-\omega^{\mathrm{el}}_R-\omega^{(\nu)}_{\mathbf{K}_\parallel})|\mathbf{u}^{(\nu)}_{\mathbf{K}_\parallel}(\mathbf{r})\cdot\mathbf{n}_z|^2$. Therefore, this analysis suggests that the light emission can be enhanced by choosing optical antennas with large ρ_{opt}.

However, we emphasize that Eq. (18) includes processes inside the metals, provided that we use the wave functions of Eqs. (2) and (3) corresponding to the QDS, whereas Bardeen's approximation considers only processes in the gap. When taking the square modulus of Eq. (18) within the QDS, we obtain the contributions of the gap, the metal electrodes and a mixed term which is a quantum interference between the two processes. Hence, it appears that it is not necessary to invoke a hot-electron mechanism to obtain a contribution of light emission from the metallic electrodes. Furthermore, we emphasize that the inelastic tunneling rate is proportional to the projected local optical density of states only within Bardeen's approximation. This is no longer necessarily true when considering the QDS, due to the interferences between the amplitudes of the SPP electric field in the gap and in the metal.

2. Energy-loss model and Poynting vector flux calculation

An alternative model [37] that has been introduced to describe light emission from tunneling junctions by Davis is to calculate the rate of energy dissipation by the electronic current:

$$\mathcal{P} = -\int_V d\mathbf{r}\,\mathbf{j}(\mathbf{r},t)\cdot\mathbf{E}(\mathbf{r},t), \quad (22)$$

where $\mathbf{E}(\mathbf{r},t)$ is the electric field generated by the current density $\mathbf{j}(\mathbf{r},t)$. This method has also been used by other authors [34,48,49]. A slightly different formulation considers the current density as a source that emits light to the far field and integrates the flux of the Poynting vector. This approach was proposed originally by Hone, Mühlschlegel, and Scalapino in Ref. [39], and, since then, it has been a popular method starting from the implementation of Laks and Mills to describe light emission from planar junctions [35], being followed by many works [6,19,50–53]. We first show the equivalence of these two points of view using the electromagnetic energy conservation in a volume V in the stationary regime:

$$\int_V d\mathbf{r}\,\mathbf{j}(\mathbf{r},t)\cdot\mathbf{E}(\mathbf{r},t) + \mathcal{P}_{\mathrm{abs}}(t) + \mathcal{P}_{\mathrm{rad}}(t) = 0, \quad (23)$$

where $\mathcal{P}_{\mathrm{rad}}$ is the flux of the Poynting vector across a surface enclosing the volume V and $\mathcal{P}_{\mathrm{abs}}$ is the power absorbed by the matter within this volume. Within the approximation of a nonlossy metal, $\mathcal{P}_{\mathrm{abs}}(t) = 0$. The radiated power is, thus, equal to the opposite of the power transferred from the tunneling current to the field. This equality establishes the equivalence between a calculation of the Poynting vector and a calculation of the power transferred from the fluctuating currents to the field. When accounting for the unavoidable metallic losses, the emitted power is then derived by multiplying the power $\mathcal{P}$ transferred to the SPPs [Eq. (22)] with the radiative efficiency η_{rad} as discussed previously.

To proceed with the evaluation of Eq. (22), we first relate the electric field to the current density using the Green's tensor $\mathbf{G}(\mathbf{r},\mathbf{r}',\omega)$ of the MIM junction as

$$\mathbf{E}(\mathbf{r},\omega) = i\omega\mu_0\int d\mathbf{r}'\mathbf{G}(\mathbf{r},\mathbf{r}',\omega)\cdot\mathbf{j}(\mathbf{r}',\omega). \quad (24)$$

The power $\mathcal{P}$ transferred from the current to the field can, thus, be written explicitly in terms of $\mathcal{S}_{j_p j_q}(\mathbf{r},\mathbf{r}',\omega)$, which is the pq element of the power cross-spectral density tensor

of the current density given by $\overline{j_p(\mathbf{r},\omega)j_q(\mathbf{r}',\omega')} = 2\pi\delta(\omega+\omega')\mathcal{S}_{j_pj_q}(\mathbf{r},\mathbf{r}',\omega)$:

$$\mathcal{P} = \int_0^\infty d\omega 2\omega\mu_0 \int d\mathbf{r} \times \int d\mathbf{r}' \sum_{p,q} \mathcal{S}_{j_pj_q}(\mathbf{r},\mathbf{r}',\omega)\mathrm{Im}[G_{p,q}(\mathbf{r},\mathbf{r}',\omega)]. \quad (25)$$

Under this form, the emitted power has the structure of the square of the intensity times an impedance proportional to $\mathrm{Im}(G)$. This point of view has been discussed in detail in Ref. [54]. In particular, the connection between the impedance and the local density of electromagnetic states has been pointed out. We recover here the point of view introduced in the dynamical Coulomb blockade where an impedance accounting for the electromagnetic environment was introduced. To proceed, it is, thus, necessary to know $\mathcal{S}_{j_pj_q}(\mathbf{r},\mathbf{r}',\omega)$ in the nonequilibrium situation of a biased junction. In most works [35,39,55], the emitted power was calculated by using the form

$$\mathcal{S}_{j_zj_z}(\mathbf{r},\mathbf{r}',\omega) \approx \frac{eI_{\mathrm{el}}}{S^2}\frac{1-\frac{\hbar\omega}{eV_B}}{1-\exp\left(\frac{\hbar\omega-eV_B}{k_BT}\right)}\delta(\mathbf{r}-\mathbf{r}') \quad (26)$$

and integrating within the gap only. In many works, the limit at $T = 0$ K was used. It predicts a linear spectrum proportional to $eV_B - \hbar\omega$. This simple form of the current density correlation was derived from a calculation of the intensity correlation in the gap and assuming a uniform and delta-correlated correlation function. This form assumed that the current density is nonzero along the z axis normal to the interfaces of the barrier. It was further assumed that the correlation function is nonzero in the gap and zero outside.

All these assumptions hindered a comparison of this model with the Fermi's golden rule result derived in the previous section. We report in Appendix B a derivation of the power cross-spectral density of the current density for a (nonequilibrium) biased junction that avoids the previous assumptions. In this derivation, we introduce a quantum field operator for the current density. As discussed before, we use either the left or the right reservoir to perform the statistical average depending on the propagation direction. This enables us to perform the calculation of the ensemble average of operators of the form $\overline{\hat{c}_1^\dagger\hat{c}_2\hat{c}_3^\dagger\hat{c}_4}$. We obtain

$$\mathcal{S}_{j_pj_q}(\mathbf{r},\mathbf{r}',\omega) = \sum_{\mathbf{k}_L}\sum_{\mathbf{k}_R} j_{p,L\to R}(\mathbf{r})j^*_{q,L\to R}(\mathbf{r}') \times 2\pi\delta(\omega-\omega_L^{\mathrm{el}}+\omega_R^{\mathrm{el}})f^L_{\mathrm{FD}}(\mathbf{k}_L)\left[1-f^R_{\mathrm{FD}}(\mathbf{k}_R)\right]. \quad (27)$$

This correlation function goes beyond the previous models: (i) It is nonzero in the metallic electrodes, and (ii) it is correlated for two points belonging each to a different electrode, in marked contrast with the assumption of a delta-correlated current. We note that, at equilibrium in a homogeneous medium, the correlation function is given by the fluctuation-dissipation theorem. Here, Eq. (27) is valid for a biased junction. In Sec. II B 3, we use this explicit form to establish the equivalence with the Fermi's golden rule result obtained in Sec. II B 1. We then discuss in detail the properties of this correlation function. We see that it has all the properties that were missing in previous models and motivated the introduction of a hot-electron mechanism [31].

3. Equivalence of the methods

In order to compute from Eq. (25) the power transferred from the electronic current to the SPPs, we restrict the Green's tensor to the contribution of the surface-plasmon modes of all branches ν and all wave vectors $\mathbf{K}_\parallel$. Using the expansion of the Green's tensor over the plasmonic modes [56,57], we obtain

$$G_{p,q}(\mathbf{r},\mathbf{r}',\omega) = \sum_{\mathbf{K}_\parallel}\sum_{\nu}\frac{c^2}{\left(\omega^{(\nu)}_{\mathbf{K}_\parallel}\right)^2-\omega^2}\frac{u^{*(\nu)}_{p,\mathbf{K}_\parallel}(\mathbf{r})u^{(\nu)}_{q,\mathbf{K}_\parallel}(\mathbf{r}')}{S}. \quad (28)$$

We then apply the limit $\lim_{\epsilon\to0}\mathrm{Im}\{1/[\alpha^2-(\omega+i\epsilon)^2]\} = \pi/(2\alpha)\delta(\omega-\alpha)$, which leads to

$$\mathrm{Im}[G_{p,q}(\mathbf{r},\mathbf{r}',\omega)] = \sum_{\mathbf{K}_\parallel}\sum_{\nu}\frac{\pi c^2}{2\omega^{(\nu)}_{\mathbf{K}_\parallel}}\frac{u^{*(\nu)}_{p,\mathbf{K}_\parallel}(\mathbf{r})u^{(\nu)}_{q,\mathbf{K}_\parallel}(\mathbf{r}')}{S}\delta(\omega-\omega^{(\nu)}_{\mathbf{K}_\parallel}). \quad (29)$$

By inserting the above forms of the Green's tensor and the current density cross-spectral density into Eq. (25), we recognize the matrix elements $\mathcal{M}^{(\nu)}_{L,R,\mathbf{K}_\parallel}$ and recover Eq. (21). This establishes the equivalence between the different models provided that the exact current density cross-spectral density given by Eq. (27) is used.

4. Current-density correlation

We now turn to a detailed study of the correlation and consider the same Al-Al_2O_3-Au system as in Fig. 2. For a metal, the typical correlation length is expected to be given by the Fermi wavelength. Surprisingly, we also unveil a long-range spatial correlation over 10 nm across the barrier. At this point, we note that the Hamiltonian does not account for electron-electron and electron-phonon interaction so that the model does not include dephasing. Nevertheless, the coherence length is expected to be larger than 10 nm at ambient temperatures. For instance, the dephasing time at ambient temperature in gold is estimated

to be in the range 10–30 fs [58]. With a Fermi velocity of 1.4×10^{-6} ms^{-1}, we estimate the phase coherence length to be on the order of 14–42 nm so that our results should be valid. In Fig. 3(a), we plot $\mathcal{S}_{j_z j_z}$ calculated with Eq. (27) in the center of the gap, for different bias potentials and in the zero temperature limit. For all considered values of V_B, the results obtained with Bardeen's approximation (dots) agree almost perfectly with those obtained with the QDS (lines). Hence, Bardeen's approximation is a very accurate approach to describe current fluctuations inside

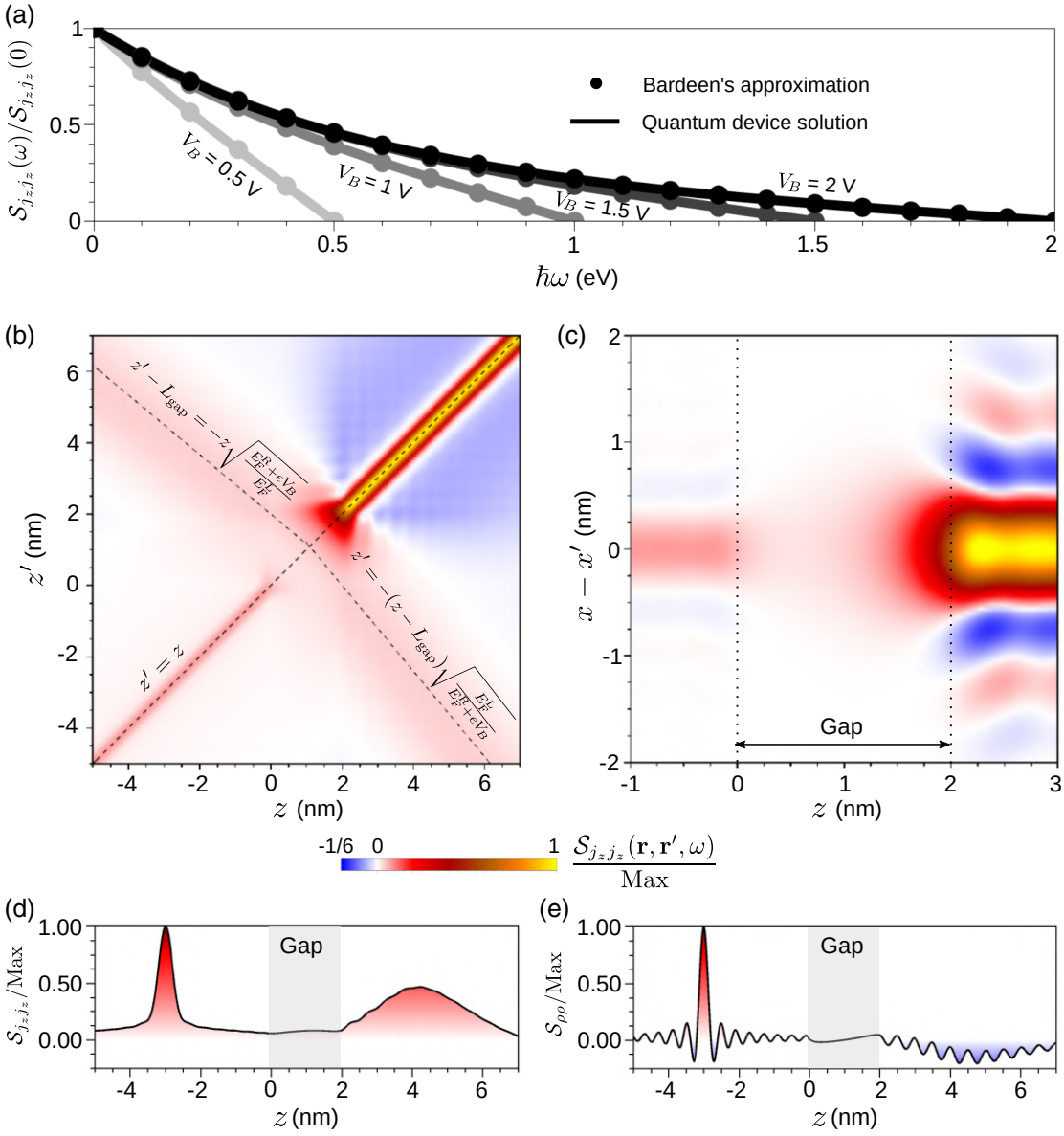

FIG. 3. Spatial correlations of the current density j_z and charge density ρ for an Al-Al_2O_3-Au tunneling junction. (a) Cross-spectral density of the current density fluctuations $\mathcal{S}_{j_z j_z}(\mathbf{r}, \mathbf{r})$ in the center of the insulator gap as a function of the energy $\hbar\omega$, for different bias potentials $V_B = 0.5$, 1, 1.5, and 2 V. Dots correspond to the calculation with the wave functions obtained within Bardeen's approximation, and the solid lines to the QDS. (b) Cross-spectral density $\mathcal{S}_{j_z j_z}(z, z')$ for points z and z' at the same position in the parallel direction ($\mathbf{r}_\parallel = \mathbf{r}'_\parallel$), for $L_{\mathrm{gap}} = 2$ nm, $V_B = 2$ V, and $\hbar\omega = 1$ eV. The insulator gap is located at values of z and z' between 0 and 2 nm. Dashed lines highlight the two peaks under the conditions $z = z'$ and $z' - L_{\mathrm{gap}} = -z\sqrt{(E_F^R + eV_B)/E_F^L}$ (and its symmetric form). (c) Cross-spectral density $\mathcal{S}_{j_z j_z}(\mathbf{r}_\parallel, \mathbf{r}'_\parallel)$ for varying positions $z = z'$ and as a function of the distance in the parallel direction $x - x'$, for the same parameters as in (b). (d),(e) Cross-spectral density of the current $\mathcal{S}_{j_z j_z}(z, z')$ and charge densities $\mathcal{S}_{\rho\rho}(z, z')$, respectively. The point $z' = -3$ nm is fixed, and the correlations are calculated for varying second point z (with $\mathbf{r}_\parallel = \mathbf{r}'_\parallel$), in the same system and for same ω as in (b) and (c).

the gap. However, we note that a linear behavior with the voltage is observed only for the smallest bias potential considered, $V_B = 0.5$ V. For larger bias potentials ($V_B = 1.5$ and 2.0 V), $\mathcal{S}_{j_z j_z}$ is not linear with respect to ω as opposed to the prediction of Eq. (26) at $T = 0$ K. Therefore, the exact definition of Eq. (27) must be used in this approach to calculate the intensity of the emitted light. This issue has been discussed in two recent papers [53,59].

We now discuss the spatial dependence of the correlation between two arbitrary points of the full device, taking advantage of the QDS to explore the currents in the metals. With this aim, we plot in Fig. 3(b) the cross-spectral density $\mathcal{S}_{j_z j_z}(z, z', \omega)$ for fixed values $L_{\text{gap}} = 2$ nm, $V_B = 2$ V, and $\hbar\omega = 1$ eV, by varying the positions z and z' (for the same position in the parallel direction, $\mathbf{r}_\parallel = \mathbf{r}'_\parallel$). We observe that there is a peak at $z = z'$ with a width of the order of 1 nm, close to the Fermi wavelengths of the metals, which are around 0.52 nm for gold and 0.36 nm for aluminum [43]. Therefore, the consideration of delta-correlated currents in Eq. (26) is accurate enough to describe this feature for most purposes. Nevertheless, we observe a second weaker and broader correlation peak for positions z and z' at opposite metals. Upon inspection (see Appendix C), this maximum occurs when the current density $\mathbf{j}_{L\to R}$ has a similar phase for all $L \to R$ transitions. The equations describing this condition are (i) $(z' - L_{\text{gap}}/z) = -(k_{zR}/k_{zL}) = -\sqrt{(E_F^R + eV_B/E_F^L)}$, if z is in the left metal and z' in the right metal, and (ii) $(z - L_{\text{gap}}/z') = -\sqrt{(E_F^R + eV_B/E_F^L)}$ if z is in the right metal and z' in the left metal. Both cases are indicated by dashed lines in Fig. 3(b).

We further analyze how the currents are correlated for different points in the direction parallel to the interfaces $\mathbf{r}_\parallel - \mathbf{r}'_\parallel \neq 0$, and for different values $z = z'$, as shown in Fig. 3(c) (we consider $x \neq x'$ but $y = y'$). The correlations are oscillatory as a function of the difference $x - x'$ and become weaker for distances larger than the Fermi wavelength. The periodicity of these oscillations is of the order of 1 nm. We also observe that the width of the peak of $\mathcal{S}_{j_z j_z}(\mathbf{r}_\parallel, \mathbf{r}'_\parallel, \omega)$ is not constant for $z = z'$ close to the insulator gap and becomes broader inside the gap. This effect can be easily included in the approximate Eq. (26) by broadening the delta function $\delta(\mathbf{r} - \mathbf{r}')$ in the parallel direction. Therefore, the assumption of delta-correlated currents in Eq. (26) needs small corrections to describe current correlations along the $\mathbf{r}_\parallel$ direction due to the finite thickness of the corresponding peak, but the assumption completely fails to capture the second peak observed for the correlations along the z direction.

In order to understand the physical origin of the current correlations in the two metals, we now focus on the correlations of the electronic charge density fluctuations. The charge density $\rho_{L\to R}$, associated to each transition from an initial state $\Psi_L^{\text{QDS}}(\mathbf{r})$ to a final state $\Psi_R^{\text{QDS}}(\mathbf{r})$, is obtained from the current density $\mathbf{j}_{L\to R}$ of Eq. (19) by using the continuity equation

$$\nabla \cdot \mathbf{j}_{L\to R} + \frac{\partial \rho_{L\to R}}{\partial t} = 0. \tag{30}$$

The cross-spectral density $\mathcal{S}_{\rho\rho}$ is then calculated with Eq. (27) after substituting $\mathbf{j}_{L\to R}$ with $\rho_{L\to R}$.

For the analysis, we fix the point $z' = -3$ nm in the left metal and observe the cross-spectral density of the current density $\mathcal{S}_{j_z j_z}$ [Fig. 3(d)] and charge density $\mathcal{S}_{\rho\rho}$ [Fig. 3(e)] for any second position z and for $\mathbf{r}_\parallel = \mathbf{r}'_\parallel$. The latter correlations oscillate more strongly in space as compared to those associated to the current density. In the results of $\mathcal{S}_{\rho\rho}$, we observe a clear peak in the position $z' = z$ as happens for $\mathcal{S}_{j_z j_z}$. More importantly, the second peak in the opposite metal also appears, but, whereas the correlation is positive for the current density, we obtain negative values in the case of the charge density. We attribute this result to the presence of a hole with positive charge and opposite velocity. In other words, as a negative charge moves toward the gap in the left electrode, a positive charge also moves toward the barrier in the right electrode. The double peak in $\mathcal{S}_{\rho\rho}$ and $\mathcal{S}_{j_z j_z}$ disappears inside the gap, where the electron and the hole recombine. These currents in the metal electrodes give a contribution to the light radiation apart from the correlations just inside the gap. This physical phenomenon had not been accounted for so far. It is included in the full QDS and introduces in a natural way a contribution of the currents in the metallic electrodes to the plasmon excitation without the need to invoke hot-electron mechanisms.

5. Discussion

In closing this theoretical section, we discuss the validity conditions of the model and its possible extensions. The discussion has been limited to planar surfaces and to a linear expansion through the use of the Fermi's golden rule excluding nonlinear effects such as two-photon emission. We stress that the equivalence between the quantum formalism and the computation of the field radiated by current fluctuations holds only within this approximation. We now discuss the extension of the model to provide guidance to further work.

Superbunching of photons attributed to two-photon spontaneous emission has been observed for a highly localized plasmonic mode between a tip and a surface [60]. This effect cannot be explained by computing the field radiated by fluctuating currents. By contrast, Muniz *et al.* [61] have shown how to account for two-photon emission using an interaction Hamiltonian with a quantized field and going to second order. Since our framework is based on a quantized form of the plasmon field, the approach used by Muniz *et al.* [61] could be implemented. It is expected that the two-photon emission contribution should increase

significantly for highly confined fields in the so-called picocavity regime.

We now discuss how to extend the model discussed in this paper, which is restricted to planar surfaces, to more general systems. Under the linear approximation, the emission can be described using the Green's tensor and the current density correlation function. Interestingly, the transverse correlation length is found to be on the order of 1 nm, confirming the assumption introduced by Laks and Mills [35]. As a consequence, the emission process is highly localized. Hence, the correlation model derived in the plane-plane geometry could be used locally provided that the local radius of curvature is larger than the transverse coherence length. Computing the emission can then be done by calculating the Green's tensor for the actual geometry and integrating over the surface. This approach has been used in the past to compute light emission by gratings [19], where it was assumed that the current density was delta-correlated in the transverse plane.

In summary, we have revisited the theoretical description of light emission in tunneling junctions by using the interaction Hamiltonian $\hat{H}_{\text{el-SPP}}$ of electrons and SPPs with a quantized form of the electric field of the plasmon mode. We have used the quantum field operator form of the current density to derive its correlation function. With this new formalism, we have established in Sec. II B the equivalence between the model based on Fermi's golden rule and the model based on the energy dissipation by a current in the linear approximation. We have found that light is emitted both in the gap and in the electrodes. In practice, these two contributions depend on the amplitude of the electric field of the plasmon modes in the gap and in the electrodes. We explore in the next section how significant are these contributions for the different plasmonic modes of the junction.

III. LIGHT EMISSION FROM PLANAR MIM JUNCTIONS

In this section, we analyze the contribution to light emission from the different plasmon modes of a device composed by metallic electrodes of approximately 10–20 nm thickness at both sides of the gap. The slow mode of the MIM junctions is localized in the insulator gap. Its electric field is strongly confined, and it is, therefore, very large so that it couples efficiently to the tunneling electrons [2,6,38,62]. On the other hand, this mode is characterized by a small group velocity and has no radiative losses unless the surface becomes rough or the metal thickness is small enough to allow for radiative leakage; thus, it is possible to engineer these radiative losses [19]. Other SPP modes are localized at metal-dielectric interfaces a few nanometers far from the insulator gap and can also contribute to radiation [3,31,53]. Their coupling to the inelastic tunneling mechanism is less efficient, especially at junctions with thick metallic electrodes, because the electric field of the SPP of the corresponding interfaces penetrates weakly into the gap. On the other hand, their radiative losses can be larger. With the objective to analyze the contribution to inelastic emission from the gap and from the metal electrodes, we choose as a representative system a junction that was considered in the first experiment of light emission from tunneling junctions, consisting in a planar junction formed by aluminum and gold electrodes separated by a layer of aluminum oxide.

A. Structure of plasmonic modes of the system

In order to study the intensity of light emission due to the process of inelastic tunneling in Al-Al_2O_3-Au junctions (where electrons tunnel from the Al electrode to the Au electrode), we first analyze the properties of the SPPs of the system. These modes are obtained by assuming that the insulators have a nondispersive permittivity, with $\varepsilon = 3.1$ for aluminum oxide in the gap. Furthermore, the aluminum layer is deposited over a glass substrate with a representative permittivity $\varepsilon = 2.5$, and the insulator in the opposite direction is set to be vacuum. The two metals are represented by a Drude permittivity of the form $\varepsilon = \varepsilon_\infty - \omega_p^2/\omega^2$, with parameters $\omega_p = 14.7\,\text{eV}$ and $\varepsilon_\infty = 1$ for aluminum [63], whereas we choose $\omega_p = 9.065$ eV for gold [64]. Furthermore, we consider interband transitions in gold by setting $\varepsilon_\infty = 9$. The losses in the Drude model are neglected in the calculation of the SPP excitation rates to ensure that the energies of the modes are real and their respective electric fields can be normalized following the quantization rule from Eq. (15). We have checked that the dispersion relations in the range of energies considered change only slightly after including losses in the permittivities. Furthermore, we fix the metal thickness of the aluminum layer at $L_{\text{Al}} = 10$ nm and that of the gold layer at $L_{\text{Au}} = 20$ nm (unless stated otherwise). Since these two thicknesses are on the order of the electron mean free paths of their respective metal [65], we assume that the electronic wave functions given in Sec. II A are valid in the whole metallic regions. A sketch of the system is shown in the inset in Fig. 4(a).

This system contains three different modes of SPPs, typically referred to as the fast, intermediate velocity, and slow modes, based on their group velocities according to the dispersion relations [shown in Fig. 4(a)] [66]. The group velocities of the fast and the intermediate velocity modes are very close to the speed of light in vacuum and in the glass, respectively, while it is much smaller for the slow mode. Among these three modes, the slow mode [brown line in Fig. 4(a)] has received significant attention in studies of quantum tunneling due to its large electric field in the gap region. Figure 4(b) illustrates the large sensitivity of the group velocity on the gap thickness. The normalized electric field distribution $\mathbf{u}_{\mathbf{K}_\parallel}^{(s)}$ of the slow mode at energy $\hbar\omega_{\mathbf{K}_\parallel}^{(s)} = 2$ eV and thickness $L_{\text{gap}} = 3$ nm is shown in

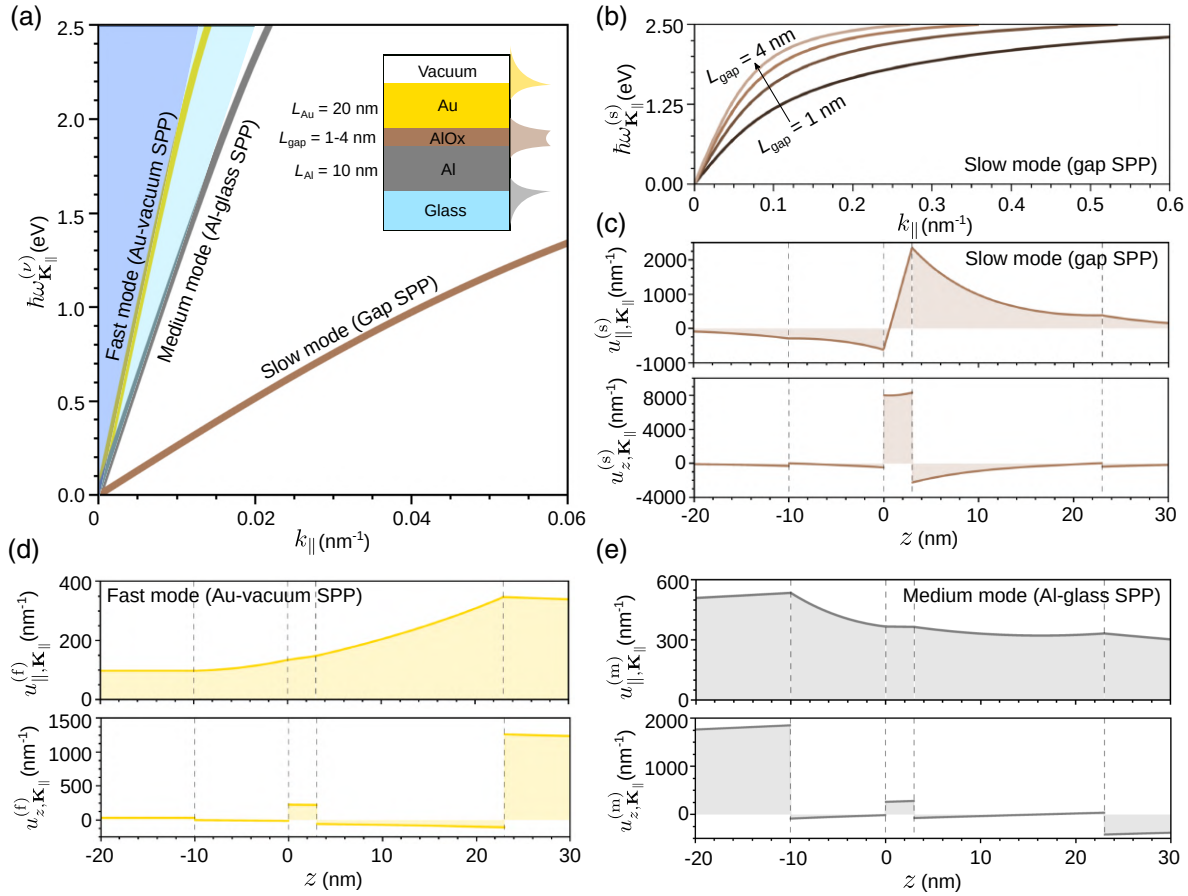

FIG. 4. SPPs in Al-Al_2O_3-Au tunneling junctions. (a) Dispersion relation of the fast (yellow), intermediate velocity (gray), and slow (brown) plasmonic modes for $L_{gap} = 3$ nm. The dark blue area indicates the light cone in vacuum, and the light blue area highlights the light cone in the glass substrate. The inset shows a sketch of the device, including the thicknesses of the layers considered in the rest of this figure and in most of the calculations throughout the manuscript. On the right of the structure, we show schematically the field distribution of the different SPP modes, emphasizing the interface where each of them propagates predominantly. Each field distribution is plotted in the same color as the corresponding dispersion (yellow for the fast mode, gray for the intermediate velocity mode, and brown for the slow mode). (b) Dispersion relation of the slow mode for different gap thicknesses $L_{gap} = 1$, 2, 3, and 4 nm (from the darkest to the lightest brown). (c)–(e) Electric field distributions at energy $\hbar\omega^{(\nu)}_{\mathbf{K}_\parallel} = 2$ eV of the slow (c), fast (d), and intermediate velocity (e) modes, for $L_{gap} = 3$ nm. The top shows the distribution of the electric field component oriented along the direction parallel to the interfaces of the junction $u^{(\nu)}_{\parallel,\mathbf{K}_\parallel}$, and the bottom shows the component in the z direction $u^{(\nu)}_{z,\mathbf{K}_\parallel}$. Dashed lines in (c)–(e) indicate the positions of the metal-insulator interfaces.

Fig. 4(c), where we plot the field components in the directions (top) parallel and (bottom) perpendicular to the interfaces. Since we neglect losses in the permittivities of the metals, these fields are real (or purely imaginary) in the whole space. It is apparent that the slow mode is strongly confined in the gap. However, it is worth noting that the electric fields also penetrate in the metals, which can contribute to the coupling of the electric field with the electronic alternating current and modify the excitation rate when considering processes in metals within the framework of the QDS, as discussed below.

Together with the slow mode, the finite thicknesses of the metals lead to the existence of two additional modes localized at the other two metal-insulator interfaces. The fast and intermediate velocity modes are SPPs mostly localized at the gold-vacuum and aluminum-substrate interfaces, respectively. These modes follow the typical dispersion relations of SPPs calculated with semi-infinite Drude metals [67], and the materials at a distance of a few nanometers from these interfaces cause only slight modifications to the dispersion relations.

The dispersion relation [Figs. 4(a) and 4(b)] and the field distribution of the modes [Figs. 4(c)–4(e)] are useful to analyze which mode may contribute to light emission. The field distributions of the fast and intermediate velocity modes suggest a smaller electromagnetic energy stored in

the gap compared to the slow mode (shown for $\hbar\omega_{\mathbf{K}_\parallel}^{(s)} = \hbar\omega_{\mathbf{K}_\parallel}^{(m)} = \hbar\omega_{\mathbf{K}_\parallel}^{(f)} = 2$ eV and $L_{\mathrm{gap}} = 3$ nm), which results in a smaller excitation rate due to a weaker coupling with tunneling electrons. However, despite its small excitation rate, only the fast mode may contribute to light emission through leakage in the glass, because it is the only mode whose dispersion relation is inside the light cone of the glass [Fig. 4(a)]. On the other hand, if the gold-vacuum interface is rough [39] or periodically structured [19], the slow mode can be coupled into vacuum, and it can dominate the light emission process due to its far larger excitation efficiency. In the following subsections, we analyze the effect of the electric field distributions on the theoretical prediction of light emission from the gap and the metal regions for the fast and slow modes. For completeness, the analysis of the intermediate velocity mode is included in Appendix D.

B. Excitation rate of the slow mode

In this section, we explore the excitation rate of the slow mode. Within Bardeen's approximation, this calculation amounts to computing an overlap integral between the current density and the plasmon field in the gap [i.e., the integral in Eq. (18) is calculated only inside the gap volume V_{gap}]. Within the QDS model, we also need to explore the contribution to the excitation rate from the processes in the metal regions and from the quantum interferences between the gap and the metal contributions.

1. Gap contribution to the slow mode excitation

We start by using Bardeen's approximation and the formalism of inelastic tunneling (described in Sec. II B) to calculate the excitation rate $\Gamma_{\mathrm{inel}}^{(s)}$ of the slow mode [from Eq. (20)]. We plot $\Gamma_{\mathrm{inel}}^{(s)}$ (where the superscript s refers to the slow mode) in the inset in Fig. 5(a) as a function of the bias potential for a fixed gap thickness of $L_{\mathrm{gap}} = 3$ nm. As V_B increases, the transition rate grows due to an exponential increase of the matrix element and a linear raise of the number of initial and final states. For instance, increasing the bias potential from 0.6 to 2.4 V causes the number of excited SPPs to increase by 3 orders of magnitude, from 3.8×10^{18} to 1.2×10^{21} SPPs per second and square meter. However, the elastic tunneling rate also increases significantly with V_B, which means that the efficiency of the tunneling junctions, corresponding to the number of excited slow SPPs per tunneling electron, raises only slightly from 10^{-4} to 2×10^{-4} for the range of V_B considered. We also check numerically that the efficiency of the junction improves for thinner gaps because the density of states of the slow mode is larger [as can be deduced from the dispersion relations shown in Fig. 4(b)]. For example, at $L_{\mathrm{gap}} = 1$ nm, we obtain an efficiency of around 8×10^{-4} SPPs excited per tunneling electron. Therefore, the efficiency of the planar junctions according to the inelastic tunneling process is not expected to exceed the ratio $\Gamma_{\mathrm{inel}}^{(s)}/\Gamma_{\mathrm{el}} = 10^{-3}$, even for narrower gaps that are experimentally considered in typical light-emission experiments with planar junctions.

2. Metal contribution to the slow mode excitation

To determine whether the inelastic tunneling in the gap can fully account for the excitation of the slow mode, we include the contribution of the metal electrodes according to the QDS model, as explained in Sec. II B. After analyzing the SPP excitation rate $\Gamma_{\mathrm{inel}}^{(s)}$ in the inset in Fig. 5(a), we now focus on the power $\mathcal{P}^{(s)}$ transferred by the current to excite this mode [given by Eq. (21)]. With this purpose, we show in Fig. 5(a) the spectral contribution $P^{(s)}(\hbar\omega_{\mathbf{K}_\parallel}^{(s)})$ at each energy $\hbar\omega_{\mathbf{K}_\parallel}^{(s)}$, which is related to the total nonradiative power $\mathcal{P}^{(s)}$ and to the slow SPP-excitation rate $\Gamma_{\mathrm{inel}}^{(s)}$ as $\mathcal{P}^{(s)} = \int P^{(s)}(\hbar\omega_{\mathbf{K}_\parallel}^{(s)}) d(\hbar\omega_{\mathbf{K}_\parallel}^{(s)}) = \int \hbar\omega_{\mathbf{K}_\parallel}^{(s)} [d\Gamma_{\mathrm{inel}}^{(s)}/d(\hbar\omega_{\mathbf{K}_\parallel}^{(s)})] d(\hbar\omega_{\mathbf{K}_\parallel}^{(s)})$. We use the electronic states of the QDS [Eqs. (2) and (3)] to calculate the spectral power $P_{\mathrm{QDS}}^{(s)}$ according to the processes in the whole MIM device [performing the integral of Eq. (18) in the volumes V_{gap} of the gap and V_{met} of the metals], and we compare it with the contribution of the gap according to Bardeen's approximation, $P_{\mathrm{BA}}^{(s)}$ [performing the integral of Eq. (18) just in V_{gap}].

For bias potentials $V_B = 0.6$, 1.2, or 1.8 V, the results of the full QDS (solid line) are nearly identical to those from Bardeen's approximation (dots) for all energies, with a largest mismatch of 5% in the integrated nonradiative power $\mathcal{P}^{(s)}$. Indeed, for all these values of V_B, the negative permittivity of the metals is large for all energies $\hbar\omega_{\mathbf{K}_\parallel}^{(s)} \leq eV_B$, resulting in limited penetration of the electric field within these regions. Furthermore, at energies $\hbar\omega_{\mathbf{K}_\parallel}^{(s)} > eV_B$, the spectral power vanishes completely, because we assume zero temperature in all this paper and the Fermi-Dirac occupation factors of the metals do not allow any transition between states at those energies (for $T > 0$ K, it is possible to excite SPPs at $\hbar\omega_{\mathbf{K}_\parallel}^{(s)} > eV_B$ [68], but we set $T = 0$ K for simplicity, because the main results of this work remain very similar otherwise). Since the calculation within Bardeen's approximation agrees with high accuracy with the calculation of the full QDS at all $\hbar\omega_{\mathbf{K}_\parallel}^{(s)}$, one can conclude that the consideration of the inelastic processes just in the insulator gap would be accurate enough to describe the excitation of the slow SPP in the range of V_B considered.

For $V_B = 2.4$ V, both calculations still agree with high accuracy at low energies, but one can observe differences

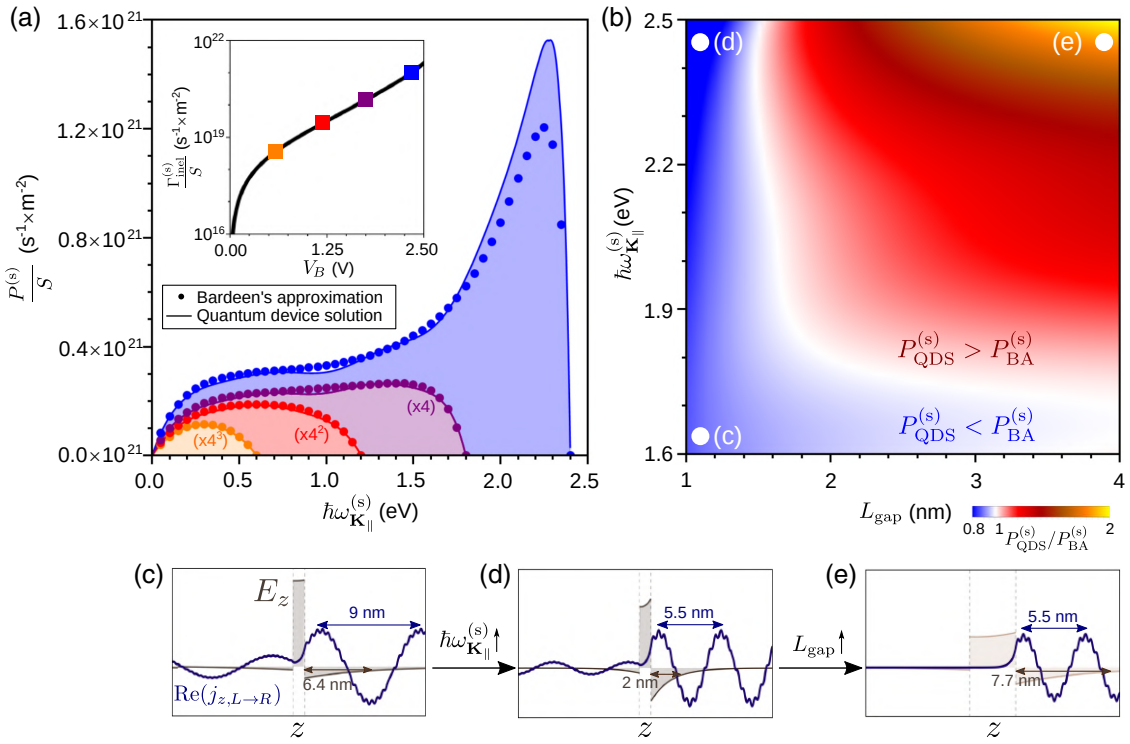

FIG. 5. Analysis of the power transferred by the tunneling current to the slow mode. (a) Spectral power $P^{(s)}(\hbar\omega^{(s)}_{\mathbf{K}_\parallel})$ per surface area S of the junction, with thicknesses $L_{\rm gap} = 3$ nm, $L_{\rm Al} = 10$ nm, and $L_{\rm Au} = 20$ nm [see the sketch in Fig. 4(a)] and bias potentials $V_B = 0.6$ (orange), 1.2 (red), 1.8 (purple), and 2.4 V (blue). Dots correspond to the gap contribution to the power, $P^{(s)}_{\rm BA}(\hbar\omega^{(s)}_{\mathbf{K}_\parallel})$, according to Bardeen's approximation, and the solid lines to the joint contribution of processes in the gap and in the metals calculated using the QDS, $P^{(s)}_{\rm QDS}(\hbar\omega^{(s)}_{\mathbf{K}_\parallel})$. The inset shows the total inelastic tunneling rate $\Gamma^{(s)}_{\rm inel}$ per surface area in logarithmic scale as a function of the bias potential, obtained within Bardeen's approximation. Colored squares correspond to the values of V_B that we choose in the main figure. (b) Ratio between the spectral power contributions obtained within the QDS and Bardeen's approximation, $P^{(s)}_{\rm QDS}(\hbar\omega^{(s)}_{\mathbf{K}_\parallel})/P^{(s)}_{\rm BA}(\hbar\omega^{(s)}_{\mathbf{K}_\parallel})$, as a function of the gap thickness $L_{\rm gap}$ and SPP energy $\hbar\omega^{(s)}_{\mathbf{K}_\parallel}$ of the Al-Al_2O_3-Au planar system. The applied bias voltage is $V_B = 3$ V. The color bar is in linear scale. (c)–(e) Distributions along the z direction of the fluctuating electronic current density $\mathrm{Re}(j_{z,L\to R})$ for an electron initially in the highest occupied energy level (blue) and of the electric field E_z (brown) of the slow mode. In each panel, we consider the gap thickness $L_{\rm gap}$ and energy $\hbar\omega^{(s)}_{\mathbf{K}_\parallel}$ indicated by each of the dots in (b): (c) $\hbar\omega^{(s)}_{\mathbf{K}_\parallel} = 1.6$ eV and $L_{\rm gap} = 1$ nm; (d) $\hbar\omega^{(s)}_{\mathbf{K}_\parallel} = 2.5$ eV and $L_{\rm gap} = 1$ nm; and (e) $\hbar\omega^{(s)}_{\mathbf{K}_\parallel} = 2.5$ eV and $L_{\rm gap} = 4$ nm. The wavelength of the electronic current density and the decay length of the SPP are indicated in each panel.

for $\hbar\omega^{(s)}_{\mathbf{K}_\parallel} \gtrsim 1.8$ eV. In this region, $P^{(s)}(\hbar\omega^{(s)}_{\mathbf{K}_\parallel})$ is dominated by a peak, where the plasmonic density of states increases considerably [see the dispersion relation in Fig. 4(b)]. At this peak, the spectral power according to the calculation of the QDS, $P^{(s)}_{\rm QDS}$, is generally larger than the value obtained within Bardeen's approximation, $P^{(s)}_{\rm BA}$, suggesting that the metal contribution gains importance under these conditions.

To further showcase the importance of the metal contribution in the high-energy regime, we plot in Fig. 5(b) the ratio $P^{(s)}_{\rm QDS}/P^{(s)}_{\rm BA}$ for varying energies and gap thicknesses under a larger bias potential of $V_B = 3$ V. Two distinct regions are observed: one at thin gaps or low energies, where the contribution within the metals reduces the excitation power of the slow mode ($P^{(s)}_{\rm QDS} < P^{(s)}_{\rm BA}$), and another at thick gaps and high energies, where it increases ($P^{(s)}_{\rm QDS} > P^{(s)}_{\rm BA}$). To clarify this phenomenon, the spatial distribution of the electric field of the slow mode is shown in Figs. 5(c)–5(e) (brown line and background), together with the inelastic current $\mathrm{Re}(j_{z,L\to R})$ associated with a $L \to R$ transition for an electron initially at the highest occupied energy level of aluminum (blue lines), for values of $L_{\rm gap}$ and $\hbar\omega^{(s)}_{\mathbf{K}_\parallel}$ indicated by white dots in Fig. 5(b). For thin gaps [Figs. 5(c) and 5(d) for $L_{\rm gap} = 1$ nm], the electric field is highly concentrated inside the gap, with some penetration into the metals near the gap. Because of the phase difference of the electric field between the insulator and the metal, the metal contributions tend to interfere

destructively with the gap contribution. Accordingly, the calculation of the QDS in the full device predicts a smaller excitation rate than Bardeen's approximation, as shown by the region of $P^{(s)}_{\mathrm{QDS}} < P^{(s)}_{\mathrm{BA}}$ in Fig. 5(b). The effect is more significant at large energies, where the negative electric field in the metal becomes even more concentrated close to the gap, leading to a stronger destructive interference [as can be observed by comparing Fig. 5(d) for $\hbar\omega^{(s)}_{\mathbf{K}_\parallel} = 2.5$ eV with Fig. 5(c) for $\hbar\omega^{(s)}_{\mathbf{K}_\parallel} = 1.6$ eV].

On the other hand, the inelastic electronic current $\mathbf{j}_{L\to R}$ oscillates in space. For larger L_{gap}, these oscillations can give not only destructive interferences between the processes in the gap and the metals, but also constructive ones under adequate circumstances. Because of the oscillations of $\mathbf{j}_{L\to R}$ [whose wavelength varies between 5 and 9 nm in the energy range considered in Fig. 5(b)], the integrand of the matrix element of Eq. (18) has the same sign in some regions of the metal as in the gap, leading to a constructive interference. Since the contribution close to the gap leads to a destructive interference, the wavelength of the electronic current should be small compared to the SPP decay length to have an overall constructive interference, which happens for large L_{gap} [Fig. 5(e)], because the SPP decay length increases with the gap thickness. After accounting for all constructive and destructive interferences within the metal according to the QDS, the interference averages to be constructive for all electrons in junctions with $L_{\mathrm{gap}} \gtrsim 1.6$ nm, and the excitation power at $\hbar\omega^{(s)}_{\mathbf{K}_\parallel} = 2.5$ eV can be even twice as high as that predicted by Bardeen's approximation. At larger energies, the slow mode contains significant losses, and, thus, the description based on nonlossy Hermitian Hamiltonians that we present in this work loses its accuracy.

In general, Fig. 5(b) illustrates that Bardeen's approximation can underestimate or overestimate the slow mode's excitation power up to a factor of 2. Importantly, while the slow mode is nonradiative in perfectly planar junctions, it can dominate light emission in other systems, such as in localized gap tunneling junctions [33], commonly used in STM, or in planar junctions with sufficient surface roughness. In these systems, the QDS gives a more appropriate description of radiation than Bardeen's approximation that considers only the gap contribution.

C. Excitation rate of the fast mode

The formalism of inelastic tunneling predicts that the excitation of the fast mode is far less efficient than that of the slow mode, due to the considerably weaker electric field inside the gap for the former [see Fig. 4(d)]. In particular, as shown in Fig. 6(a), the spectral power $P^{(f)}_{\mathrm{BA}}$ of the fast mode at $V_B = 0.6$ V is of the order of $10^{11}\ \mathrm{s^{-1}\,m^{-2}}$, which is 10^7 times smaller than $P^{(s)}_{\mathrm{BA}}$ for the slow mode. However, the study of the excitation rate of the fast mode is important, because in perfectly planar junctions it is the only process that leads to radiation. It was estimated that, in gratings with a periodicity of hundreds of nanometers, the emission of the fast mode may overcome the emission by the slow mode by a factor of 10^2 [3]. Indeed, the main discrepancy between the theory of inelastic tunneling and experiments was first observed for gratings [4,5]. The results that we present for planar junctions can be generalized to describe the contribution of Bardeen's approximation and the QDS in other structured tunneling junctions where the fast mode can be the leading mechanism of light emission.

1. Metal contribution to the fast mode excitation

Interestingly, when the QDS is applied, we already observe a difference from Bardeen's approximation at $V_B = 0.6$ V in the calculation of the spectra of the radiative power, $P^{(f)}$. Although this variation is only 18% in the integrated power $\mathcal{P}^{(f)}$, it is considerably larger than for the slow mode at the same bias potential. This suggests that including the metal contribution is significant for the fast mode as pointed out by Kirtley *et al.* [31]. Note that the metal contribution is obtained without using the hot-electron mechanism. The discrepancy between Bardeen's approximation and the QDS increases considerably for $V_B = 1.2$ V [Fig. 6(b)] and continues to grow with V_B. Furthermore, at $V_B = 1.8$ V [Fig. 6(c)], local minima and maxima in $P^{(f)}_{\mathrm{QDS}}$ are observed at different energies, which is related to the oscillatory behavior of the inelastic current density for each electron, as shown in the inset in Fig. 6(c) together with the electric field of the fast mode. The wavelength of the oscillations inside the metal depends on the energy of the electrons, leading to constructive or destructive interference with the contribution of the insulator gap in the integral of Eq. (18), which relates the current density according to Eq. (19) with the field distribution of the mode [Eq. (14)]. However, we do not expect these oscillations to be as predominant in experiments, because the position of the maxima and minima in the spectra $P^{(f)}_{\mathrm{QDS}}$ is very sensitive to the thicknesses of the metals and the gap. In real systems, metallic surfaces present small roughness, and, thus, the contribution of different thicknesses cancels out these oscillations and the measured power would be the average between thicknesses. Finally, we remind that the calculation of the matrix element uses a very simple model of electronic wave functions which may produce artifacts. Specifically, to compute the overlap integral between the electronic wave function and the plasmon mode, we use the electronic wave function computed for an infinite metal and perform the integral over a finite thickness.

The increased importance of the metal contribution to light emission at larger energies becomes more evident in Fig. 6(d). The comparison between the results obtained

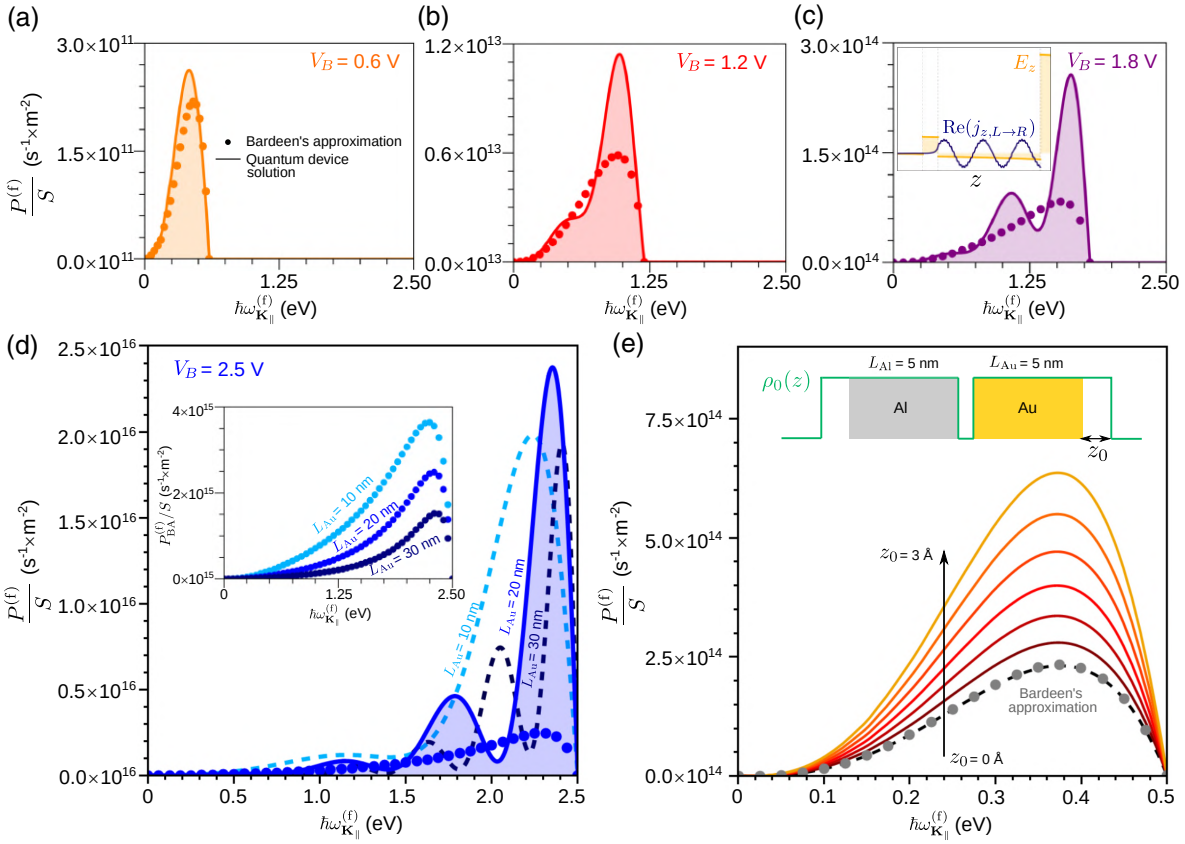

FIG. 6. Analysis of the radiative power due to the excitation of the fast mode. (a)–(c) Spectral radiative power $P^{(f)}(\hbar\omega^{(f)}_{\mathbf{K}_\parallel})$ per surface area S of the junction, with thicknesses $L_{\mathrm{gap}} = 3$ nm, $L_{\mathrm{Al}} = 10$ nm, and $L_{\mathrm{Au}} = 20$ nm [sketch in Fig. 4(a)] and bias potentials $V_B = 0.6$ (a), 1.2 (b), and 1.8 V (c). Dots indicate the contribution of the gap according to Bardeen's approximation, and the solid lines refer to the full results of the QDS. The inset in (c) shows the electric field of the fast mode at energy 1.8 eV and the oscillating inelastic current density of one electron along the z direction. (d) Spectral radiative power $P^{(f)}(\hbar\omega^{(f)}_{\mathbf{K}_\parallel})$ per surface area S under bias potential $V_B = 2.5$ V. We compare the results obtained using Bardeen's approximation (dots) and the QDS (solid line) for $L_{\mathrm{Au}} = 20$ nm. Dashed lines indicate the results according to the QDS for $L_{\mathrm{Au}} = 10$ nm and $L_{\mathrm{Au}} = 30$ nm. The inset shows the results within Bardeen's approximation for $L_{\mathrm{Au}} = 10$, 20, and 30 nm. (e) Spectral radiative power per surface area S including nonlocal effects (colored lines) for different distances z_0 from the boundary between the gold and vacuum permittivity (corresponding to the position of the centroid of charge) to the boundary of the ionic-positive background. z_0 is indicated in the inset and changes in steps of size 0.5 Å. We consider that the boundary between permittivities is inside the ionic background. The results obtained within the QDS without considering nonlocal effects ($z_0 = 0$) and within Bardeen's approximation are indicated by the black dashed line and by the gray dots, respectively. All the calculations in (e) are done for a junction with an aluminum and a gold layer of thicknesses $L_{\mathrm{Al}} = L_{\mathrm{Au}} = 5$ nm placed between a glass substrate and vacuum, with parameters $L_{\mathrm{gap}} = 2$ nm and $V_B = 0.5$ V.

with Bardeen's approximation (blue dots) and QDS (blue solid line) for $V_B = 2.5$ V demonstrates that, for most energies, the results of the latter calculation are significantly larger than for the former. Therefore, despite the oscillations in the calculation for the full system, the integrated power $\mathcal{P}^{(f)}_{\mathrm{QDS}}$ is notably larger than $\mathcal{P}^{(f)}_{\mathrm{BA}}$.

The role of the thickness of the metallic electrodes is also highlighted in Fig. 6(d). The inset indicates the result $P^{(f)}_{\mathrm{BA}}$ according to Bardeen's approximation for $L_{\mathrm{Au}} = 10$ nm, $L_{\mathrm{Au}} = 20$ nm, and $L_{\mathrm{Au}} = 30$ nm (the values of $L_{\mathrm{Au}} = 20$ nm are the same as in the main figure). Bardeen's approximation predicts that $P^{(f)}_{\mathrm{BA}}$ decreases significantly when increasing L_{Au}, because the overlap between the gap and the fast SPP field decreases. Therefore, the decrease of the inelastic tunneling rate is dictated by the decay length of the SPP. However, experimental measurements from Ref. [4] indicate that the intensity of the light emitted by the fast mode decreases with L_{Au} more slowly than the decay length of the SPP. To assess if the QDS can account for this experimental result, we plot in Fig. 6(d) the radiative power for the same thicknesses $L_{\mathrm{Au}} = 10$, 20, and 30 nm. The obtained radiative power decreases more slowly as a function of L_{Au} than expected from Bardeen's approximation. In general, Figs. 6(a)–6(d) show that, under different circumstances, the QDS can lead to a value of the power

$\mathcal{P}^{(f)}_{\mathrm{QDS}}$ 2 or 3 times larger than $\mathcal{P}^{(f)}_{\mathrm{BA}}$ for intermediate bias potentials or even an order of magnitude larger at high energies and thick metallic layers.

2. Nonlocal contribution to the fast mode excitation

Together with the inelastic processes inside the metallic regions, Kirtley *et al.* argued that processes at the vacuum-metal interfaces placed far from the insulator barrier could also contribute significantly to the excitation of the fast mode [31]. The importance of this interface was also shown by experiments, notably when demonstrating experimentally that the presence of adsorbants under ultrahigh vacuum could quench the signal [30]. It is seen in Fig. 4(d) that the electric field changes drastically from the bulk metal to vacuum at the interface due to dielectric screening. Therefore, one could expect that, under a nonlocal description of the electromagnetic response, the tunneling electrons could interact efficiently with an unscreened electric field of the fast SPP outside the metal over a distance given by the screening length.

To estimate the possibility of plasmon excitation at the interface enhanced by nonlocal effects, we examine a system composed of an aluminum electrode and a gold electrode of thicknesses $L_{\mathrm{Al}} = L_{\mathrm{Au}} = 5$ nm separated by an insulator gap of thickness $L_{\mathrm{gap}} = 2$ nm. These values are taken from Ref. [53], where it was noted that the gap contribution alone does not explain all the light emitted by the fast mode of this system, even at low bias potentials. We show in Fig. 6(e) the spectral radiative power $P^{(f)}$ of the fast SPP of the mentioned junction for $V_B = 0.5$ V. In these circumstances, the calculation of the excitation power with the QDS $P^{(f)}_{\mathrm{QDS}}$ and in the absence of any nonlocal effect (dashed line) is very similar to the gap contribution $P^{(f)}_{\mathrm{BA}}$ according to Bardeen's approximation (dots), because (i) at such low bias potentials, the field penetration is small at all energies $\hbar\omega^{(f)}_{\mathbf{K}_\parallel} < eV_B$ due to the large negative permittivities of the metals, so that the fields at the gap are comparatively large; and (ii) at so thin metallic layers, the space to excite the fast SPP in the metals is not large compared to the gap where the electric fields are confined. However, the performed calculations assume that the electron density is nonzero only up to the interfaces between the gold and vacuum permittivities, where the limits of the integrals in Eq. (18) are set.

According to microscopic descriptions of metallic surfaces, the electronic density in real metals may surpass the position of the interface considered in classical electromagnetism [10,69–73]. We now consider that the electronic current shifts by a maximum distance z_0 with respect to the boundary between permittivities, as shown in the inset in Fig. 6(e), according to the simple model of nonlocality detailed in Appendix E. This simple model results in electrons tunneling from the first electrode and reaching positions near the metal-vacuum interface of the second electrode where the electric field is strong, which efficiently boosts the coupling to electrons. Specifically, we consider that electrons can interact with the electric field outside the junction in a region of different widths $z_0 = 0.5$, 1, 1.5, 2, 2.5, and 3 Å [from brown to yellow lines in Fig. 6(e)]. The power transferred to the fast SPP becomes considerably larger for increasing z_0. For example, for $z_0 = 3$ Å, the QDS predicts an excitation power $P^{(f)}_{\mathrm{QDS}}$ 3 times larger than $P^{(f)}_{\mathrm{BA}}$. We have, thus, shown that the QDS can account not only for processes in the gap and in the bulk metal, but also for those at the interface. The model is, thus, able to describe systems where any contribution, and not only from inelastic tunneling processes in the gap, is relevant.

IV. CONCLUSIONS

In this paper, we have introduced a theoretical approach to describe light emission from biased planar tunnel junctions. We stress that previous works mostly used Bardeen's approximation. We revisited the two models that have been used so far to describe the inelastic tunneling mechanism: a Hamiltonian description of light-matter interaction based on Fermi's golden rule and the calculation of the power radiated by fluctuating currents. The latter model requires the cross-spectral density of the fluctuating electronic current. These models were used independently in the past. Here, we derived the form of the cross-spectral density of the current density for a biased junction. We found that the electronic current density is strongly correlated in the opposite metals. These long-range correlations can be interpreted as an electron-hole pair that recombines in the gap and were missing in previous models. They provide a model that includes electronic and photonic interactions in the metallic electrodes. With this long-range correlation function, we have been able to establish explicitly the equivalence of the approach based on the Fermi golden rule and the approach based on the calculation of the power radiated by the current density fluctuations.

A detailed analysis of the effect of the long-range correlations showed that they hardly contribute to the excitation rate of the slow plasmonic mode of a planar junction. This is in good agreement with the fact that Bardeen's approximation is valid to describe light emission in planar tunneling junctions and also for STM light emission. By contrast, we have found that this correction is the major contribution for the fast mode excitation which often dominates in the case of periodically corrugated planar junctions.

Using the QDS model of light emission by inelastic tunneling, we have been able to reproduce numerically three experimental observations that could not be explained by the previous models. First, it had been reported that the existing models underestimated the measured emitted power. We found that the QDS model increases the emitted

power up to an order of magnitude. Second, we find that the QDS model predicts a decay of the power when the thickness of the metal increases slower than the exponential decay of the fast SPP. Finally, we introduced an effective model of nonlocality that accounts for a large excitation of the fast mode at the vacuum-metal interface and also for the quenching of light emission due to adsorbants. On the whole, it does no longer appear necessary to invoke an alternative process based on the existence of hot electrons.

In addition to these previously unexplained observations that motivated our analysis, the QDS opens new possibilities. Optimizing the choice of the material and thickness of the metallic layer sustaining a fast plasmon mode could lead to a large coupling due to nonlocal effects. Another very interesting direction for future work is the modeling and optimization of two-photon spontaneous emission which leads to bunching as reported recently [60]. The two-photon emission process can be computed using an interaction Hamiltonian with a quantized field and a second-order contribution [61]. Hence, the formalism introduced in this paper based on a quantized plasmon could be applied. Finally, the framework introduced in this paper could be applied to light emission by quantum cascade structures in the LED regime. So far, many systems expected to produce electroluminescence assisted by plasmonic emission are operating in the incandescence regime. The experiments have been guided by a theoretical approach based on two points of view: The plasmon is an electronic collective excitation, and its excitation is described by a deterministic injection. Here, the plasmon is treated as a quantized electromagnetic field and the injection is treated as a tunneling through a barrier separating the quantum well from a medium with a well-defined Fermi level.

In summary, we have introduced a new theoretical framework to account for light emission by inelastic tunneling. It enables one to model experimental observations that were out of reach of the available models. The new framework also appears to be a useful tool to address important issues such as two-photon emission by inelastic tunneling or spontaneous emission in electrically biased cascade structures.

ACKNOWLEDGMENTS

The authors acknowledge helpful discussions with Julien Gabelli, Cheng Zhang, and Jean-Paul Hugonin. This work has been supported by Agence Nationale de la Recherche (ANR-22-CE24-0011) and Project No. IT 1526-22 from the Basque Government for consolidated groups of the Basque University, as well as Grants No. PID2019–107432 GB-I00 and No. PID2022-139579NB-I00, both funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe."

APPENDIX A: DERIVATION OF THE ELASTIC TUNNELING RATE

In order to compute the elastic tunneling rate that leads to the intensity-voltage curves of the tunneling junction [as plotted, for example, in Fig. 2(e)], we start by focusing on Bardeen's approximation (in this appendix, we omit the superscript BA in the wave functions for brevity) to recover the result that he discovered originally in Ref. [36]. We consider the transition rate from a particular state $|\Psi_L\rangle$ of the left metal (with energy $\hbar\omega_L^{\mathrm{el}}$ and parallel component of the wave vector $\mathbf{k}_{\parallel L}$) to another state of the right metal $|\Psi_R\rangle$ (with respective values $\hbar\omega_R^{\mathrm{el}}$ and $\mathbf{k}_{\parallel R}$). This rate is calculated by Fermi's golden rule with the interaction Hamiltonian $\hat{H}_{\mathrm{el}} - \hat{H}_L$ [Eq. (1)]:

$$\Gamma_{L\to R} = \frac{2\pi}{\hbar^2}\delta(\omega_L^{\mathrm{el}} - \omega_R^{\mathrm{el}})|\langle\Psi_R|\hat{H}_{\mathrm{el}} - \hat{H}_L|\Psi_L\rangle|^2. \tag{A1}$$

We first compute the matrix element of the interaction Hamiltonian. Although the wave functions depend on the three spatial coordinates as $\Psi_{L(R)}(\mathbf{r}) = \Psi_{L(R)}(z)e^{i\mathbf{k}_{\parallel L(R)}\cdot\mathbf{r}_\parallel}$, the interaction Hamiltonian depends only on z, which enables to separate the integral into spatial coordinates:

$$\begin{aligned}
|\langle\Psi_R|\hat{H}_{\mathrm{el}} - \hat{H}_L|\Psi_L\rangle|^2 &= \left(\int \Psi_R^*(z)e^{-i\mathbf{k}_{\parallel R}\cdot\mathbf{r}_\parallel}(\hat{H}_{\mathrm{el}} - \hat{H}_L)[\Psi_L(z)e^{i\mathbf{k}_{\parallel L}\cdot\mathbf{r}_\parallel}]d\mathbf{r}\right)\left(\int \Psi_L^*(z')e^{-i\mathbf{k}_{\parallel L}\cdot\mathbf{r}'_\parallel}(\hat{H}_{\mathrm{el}} - \hat{H}_L)[\Psi_R(z')e^{i\mathbf{k}_{\parallel R}\cdot\mathbf{r}'_\parallel}]d\mathbf{r}'\right) \\
&= \left|\int \Psi_R^*(z)(\hat{H}_{\mathrm{el}} - \hat{H}_L)\Psi_L(z)dz\right|^2 \int\int e^{-i(\mathbf{k}_{\parallel R} - \mathbf{k}_{\parallel L})\cdot(\mathbf{r}_\parallel - \mathbf{r}'_\parallel)}d\mathbf{r}_\parallel d\mathbf{r}'_\parallel \\
&= \left|\int \Psi_R^*(z)(\hat{H}_{\mathrm{el}} - \hat{H}_L)\Psi_L(z)dz\right|^2 (2\pi)^2\delta(\mathbf{k}_{\parallel R} - \mathbf{k}_{\parallel L})S.
\end{aligned} \tag{A2}$$

Since the Hamiltonian $\hat{H}_L$ is equal to the complete electronic Hamiltonian $\hat{H}_{\mathrm{el}}$ in the regions of the left metal and the insulator gap, the integral over z in Eq. (A2) has to be done just in the right metal. In this region, the complete

Hamiltonian $\hat{H}_{\mathrm{el}}$ is equal to $\hat{H}_R$, and, following the procedure of Bardeen in Ref. [36], we can add the vanishing term $-\Psi_L(z)(\hat{H}_{\mathrm{el}}-\hat{H}_R)\Psi_R^*(z)$ to obtain a symmetrical form inside the integral:

$$\begin{aligned}\int_{-\infty}^{\infty}\Psi_R^*(z)(\hat{H}_{\mathrm{el}}-\hat{H}_L)\Psi_L(z)dz &= \int_{L_{\mathrm{gap}}}^{\infty}\Psi_R^*(z)(\hat{H}_{\mathrm{el}}-\hat{H}_L)\Psi_L(z)-\Psi_L(z)(\hat{H}_{\mathrm{el}}-\hat{H}_R)\Psi_R^*(z)dz\\ &= \int_{L_{\mathrm{gap}}}^{\infty}\Psi_R^*(z)(\hat{H}_{\mathrm{el}}-\hbar\omega_L^{\mathrm{el}})\Psi_L(z)-\Psi_L(z)(\hat{H}_{\mathrm{el}}-\hbar\omega_R^{\mathrm{el}})\Psi_R^*(z)dz\\ &= \int_{L_{\mathrm{gap}}}^{\infty}\Psi_R^*(z)\left(-\frac{\hbar^2\nabla^2}{2m_{\mathrm{eff}}}+U(z)\right)\Psi_L(z)-\Psi_L(z)\left(-\frac{\hbar^2\nabla^2}{2m_{\mathrm{eff}}}+U(z)\right)\Psi_R^*(z)dz\\ &= -\frac{\hbar^2}{2m_{\mathrm{eff}}}\int_{L_{\mathrm{gap}}}^{\infty}\left(\Psi_R^*(z)\frac{\partial^2\Psi_L(z)}{\partial z^2}-\Psi_L(z)\frac{\partial^2\Psi_R^*(z)}{\partial z^2}\right)dz. \end{aligned} \tag{A3}$$

Here, we take into account that the energies $\hbar\omega_L^{\mathrm{el}}$ and $\hbar\omega_R^{\mathrm{el}}$ must be equal so that Eq. (A1) leads to a nonzero value. By noticing that the functions $\Psi_L(z)$ and $\partial\Psi_L(z)/\partial z$ vanish at infinity, we solve the integral using the method of integration by parts:

$$\begin{aligned}\int\Psi_R^*(z)(\hat{H}_{\mathrm{el}}-\hat{H}_L)\Psi_L(z)dz &= -\frac{\hbar^2}{2m_{\mathrm{eff}}}\left(\Psi_R^*(z)\frac{\partial\Psi_L(z)}{\partial z}\bigg|_{L_{\mathrm{gap}}}^{\infty}-\int_{L_{\mathrm{gap}}}^{\infty}\frac{\partial\Psi_L(z)}{\partial z}\frac{\partial\Psi_R^*(z)}{\partial z}dz-\Psi_L(z)\frac{\partial\Psi_R^*(z)}{\partial z}\bigg|_{L_{\mathrm{gap}}}^{\infty}\right.\\ &\quad\left.+\int_{L_{\mathrm{gap}}}^{\infty}\frac{\partial\Psi_L(z)}{\partial z}\frac{\partial\Psi_R^*(z)}{\partial z}dz\right)\\ &= \frac{\hbar^2}{2m_{\mathrm{eff}}}\left(\Psi_R^*(z)\frac{\partial\Psi_L(z)}{\partial z}-\Psi_L(z)\frac{\partial\Psi_R^*(z)}{\partial z}\right)\bigg|_{z=L_{\mathrm{gap}}}. \end{aligned} \tag{A4}$$

Equations (A1), (A2), and (A4) lead to the following transition rate between left and right states:

$$\Gamma_{L\to R}=\frac{(2\pi)^3\hbar^2}{4m_{\mathrm{eff}}^2}S\delta(\omega_L^{\mathrm{el}}-\omega_R^{\mathrm{el}})\delta(\mathbf{k}_{\parallel R}-\mathbf{k}_{\parallel L})\left|\left(\Psi_R^*(z)\frac{\partial\Psi_L(z)}{\partial z}-\Psi_L(z)\frac{\partial\Psi_R^*(z)}{\partial z}\bigg|_{z=L_{\mathrm{gap}}}\right)\right|^2. \tag{A5}$$

According to Eq. (A5), together with the energy, the parallel component of the wave vector $\mathbf{k}_{\parallel}$ must also be conserved in the transition due to the homogeneity of the system in the $\mathbf{r}_{\parallel}$ direction. Furthermore, the electronic wave functions appear in the expression of the transition rate in the form $\mathbf{j}_{\mathrm{Bar}}(z)=(i\hbar e/2m_{\mathrm{eff}})[\Psi_R^*(z)\partial_z\Psi_L(z)-\Psi_L(z)\partial_z\Psi_R^*(z)]\mathbf{n}_z$ (where $\mathbf{n}_z$ is the unit vector in the z direction) evaluated in the boundary between the gap and the right metal. In the original work by Bardeen, this term was associated to the transition current density of elastic tunneling, due to its similar form of the probability current density $\mathbf{j}(\mathbf{r})=(i\hbar e/2m_{\mathrm{eff}})[\Psi^*(\mathbf{r})\nabla\Psi(\mathbf{r})-\Psi(\mathbf{r})\nabla\Psi^*(\mathbf{r})]$ of a quantum state.

The current density measured in an experiment is due to all possible transitions from occupied states of the left metal to unoccupied states of the right metal. Thus, the rate $\Gamma_{L\to R}$ must be summed for all these transitions. We first consider the sum over all final states, which leads to the transmission probability for each incident electron through the junction. By using the Sommerfeld model of free electrons with periodic boundary conditions to define the states in each metal, all the states in the right metal have a wave vector $\mathbf{k}_R$ associated. Considering the large number of states, the discrete sum $(1/L_zS)\sum_{\mathbf{k}_R}$ can be converted into the integral $[1/(2\pi)^3]\int d\mathbf{k}_R$, which leads to the expression

$$\begin{aligned}\Gamma_L &= \sum_{\mathbf{k}_R}\Gamma_{L\to R}=\frac{L_z}{2\pi}\frac{S}{(2\pi)^2}\int\Gamma_{L\to R}\frac{dk_{zR}}{d\omega_{\mathrm{el}}^R}d\mathbf{k}_{\parallel R}d\omega_R^{\mathrm{el}}\\ &= \frac{L_zS^2\hbar^2}{4m_{\mathrm{eff}}^2}\frac{dk_{zR}}{d\omega_R^{\mathrm{el}}}\bigg|_{\omega_R^{\mathrm{el}}=\omega_L^{\mathrm{el}}}\left|\left(\Psi^*_{R(\mathbf{k}_{\parallel\mathbf{L}},\hbar\omega_L^{\mathrm{el}})}(z)\frac{\partial\Psi_{L(\mathbf{k}_{\parallel\mathbf{L}},\hbar\omega_L^{\mathrm{el}})}(z)}{\partial z}-\Psi_{L(\mathbf{k}_{\parallel\mathbf{L}},\hbar\omega_L^{\mathrm{el}})}(z)\frac{\partial\Psi^*_{R(\mathbf{k}_{\parallel\mathbf{L}},\hbar\omega_L^{\mathrm{el}})}(z)}{\partial z}\bigg|_{z=L_{\mathrm{gap}}}\right)\right|^2. \end{aligned} \tag{A6}$$

The effect of the integral over the final states in Eq. (A6) is, thus, to impose that the left and right states have the same parallel wave vector $\mathbf{k}_{\parallel L}$ and energy $\hbar\omega_L^{\mathrm{el}}$, as expected for an elastic process. Furthermore, the term $(dk_{zR}/d\omega_{\mathrm{el}}^R) = (m_{\mathrm{eff}}/\hbar k_{zR})$ includes the density of states in the metal on the right.

The total tunneling rate is then obtained by summing Γ_L over all initial states of the left metal:

$$\Gamma_{\mathrm{el}} = \sum_{\mathbf{k}_L}^{\mathrm{occ}} \Gamma_L = \frac{L_z}{2\pi}\frac{S}{(2\pi)^2}\int_0^\infty d\omega_L^{\mathrm{el}} \int_0^{\min_{j\in\{L,R\}}\sqrt{\frac{2m_{\mathrm{eff}}}{\hbar^2}(\hbar\omega_L^{\mathrm{el}}-U_j)}} dk_{\parallel L}\left[f_{\mathrm{FD}}^L(\mathbf{k}_L)[1-f_{\mathrm{FD}}^R(\mathbf{k}_R)]\Gamma_L(\hbar\omega_L^{\mathrm{el}},k_{\parallel L})2\pi k_{\parallel L}\frac{dk_{zL}}{d\omega_L^{\mathrm{el}}}\right]. \tag{A7}$$

We notice that, for each energy $\hbar\omega_L^{\mathrm{el}}$, there are electronic states with wave vectors up to a maximal value of $|\mathbf{k}_{\parallel L(R)}| = \sqrt{(2m_{\mathrm{eff}}/\hbar^2)(\hbar\omega_L^{\mathrm{el}} - U_{L(R)})}$ in the metal on the left and on the right. Because of the conservation of $\mathbf{k}_{\parallel}$, a transition is valid only if both metals have an electronic state for a vector $\mathbf{k}_{\parallel L(R)}$. Thus, the integral has to be calculated up to the minimum between the two extremal values $|\mathbf{k}_{\parallel L(R)}|$ that accept a state in both metals. Furthermore, in Eq. (A7) we impose that the initial state of wave vector $\mathbf{k}_L$ must be occupied [with probability given by the Fermi-Dirac occupation factor $f_{\mathrm{FD}}^L(\mathbf{k}_L)$] and that the final state of wave vector $\mathbf{k}_R$ must be unoccupied [with probability $1 - f_{\mathrm{FD}}^R(\mathbf{k}_R)$]. Under Bardeen's approximation, the tunneling rate per electron is obtained by replacing the wave functions from Eqs. (6) and (7) into Eq. (A6), which gives Eq. (9). Then, the elastic tunneling rate plotted in Fig. 2 (black dots) is obtained by performing the integral of Eq. (A7) with the expression of Γ_L of Eq. (9).

Alternatively, the elastic tunneling rate can be calculated using the QDS. By solving the Schrödinger equation with the Hamiltonian $\hat{H}_{\mathrm{el}}$, the wave functions follow the expressions of Eqs. (2) and (3). For the states of the left metal, the coefficients are

$$r_L = -\frac{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}, \tag{A8}$$

$$\alpha_L = -\frac{2ik_{zL}(k_{z\mathrm{gap}}+ik_R)}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}, \tag{A9}$$

$$\beta_L = -\frac{2ik_{zL}(k_{z\mathrm{gap}}-ik_R)e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}, \tag{A10}$$

$$t_L = -\frac{4ie^{k_{z\mathrm{gap}}L_{\mathrm{gap}}}e^{-ik_RL_{\mathrm{gap}}}k_{z\mathrm{gap}}k_{zL}}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}; \tag{A11}$$

and for the right metal

$$r_R = -\frac{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}+ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}-ik_{zR})}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}, \tag{A12}$$

$$\alpha_R = -\frac{2ik_{zR}(k_{z\mathrm{gap}}+ik_L)}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}, \tag{A13}$$

$$\beta_R = -\frac{2ik_{zR}(k_{z\mathrm{gap}}-ik_L)e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}, \tag{A14}$$

$$t_R = -\frac{4ie^{k_{z\mathrm{gap}}L_{\mathrm{gap}}}e^{-ik_LL_{\mathrm{gap}}}k_{z\mathrm{gap}}k_{zR}}{e^{2k_{z\mathrm{gap}}L_{\mathrm{gap}}}(k_{z\mathrm{gap}}-ik_{zL})(k_{z\mathrm{gap}}-ik_{zR}) - (k_{z\mathrm{gap}}+ik_{zL})(k_{z\mathrm{gap}}+ik_{zR})}. \tag{A15}$$

In this approach, we can directly evaluate the probability current density associated to each electronic state as $\mathbf{j}(z) = -(1/L_zS)(\hbar e k_{zR}/m_{\mathrm{eff}})|t_L|^2\mathbf{n}_z$. This value is related with the tunneling rate per electron using Eq. (11). In order

to calculate the total transition rate, we repeat the procedure of Bardeen's approximation and obtain Eq. (12). Therefore, the results plotted in Fig. 2 for the QDS (orange line) are obtained from the integral of Eq. (A7) using the corresponding expression for Γ_L [Eq. (12)].

APPENDIX B: DERIVATION OF THE CURRENT DENSITY CROSS-SPECTRAL DENSITY

We start by introducing the form of the field operator for an arbitrary number of electrons in the conduction band [74]:

$$\hat{\Psi}(\mathbf{r},t)=\sum_n \psi_n(\mathbf{r})\hat{c}_n(t). \tag{B1}$$

Here, the subscript $n=(\mathbf{k},l)$, where $l=L,R$ labels the state with wave vector $\mathbf{k}$ and propagation coming from the left or the right metal. $\hat{c}_n$ is the Fermionic annihilation operator of the n state. The current density is then given by [74]

$$\hat{\mathbf{j}}(\mathbf{r},t)=\frac{i\hbar e}{2m_{\mathrm{eff}}}[\hat{\Psi}^\dagger(\mathbf{r},t)\nabla\hat{\Psi}(\mathbf{r},t)-\hat{\Psi}(\mathbf{r},t)\nabla\hat{\Psi}^\dagger(\mathbf{r},t)]. \tag{B2}$$

The time dependence of each Fermionic operator can be cast in the form

$$\hat{c}_n(t)=\hat{c}_n\exp(-i\omega_n t), \tag{B3}$$

where $\hbar\omega_n=(\hbar^2\mathbf{k}^2/2m_{\mathrm{eff}})+U_{L(R)}$. It follows that the current density in frequency domain is

$$\hat{\mathbf{j}}(\mathbf{r},\omega)=\sum_{n,n'}\mathbf{j}(\mathbf{r},n,n')2\pi\delta(\omega-\omega_{nn'})\hat{c}_n^\dagger\hat{c}_{n'}, \tag{B4}$$

where $\mathbf{j}(\mathbf{r},n,n')=(i\hbar e/2m_{\mathrm{eff}})[\psi_n^*(\mathbf{r})\nabla\psi_{n'}(\mathbf{r})-\psi_{n'}(\mathbf{r})\nabla\psi_n^*(\mathbf{r})]$ and $\omega_{nn'}=\omega_n-\omega_{n'}$.

We can now compute the ensemble average of the current density $\overline{\hat{j}_p(\mathbf{r},\omega)\hat{j}_q(\mathbf{r}',\omega')}$, where p and q stand for the Cartesian components of the current density. To proceed, we assume that the system is a statistical mixture and use the commutation relations satisfied by Fermionic operators $\{\hat{c}_n,\hat{c}_{n'}\}=0$; $\{\hat{c}_n^\dagger,\hat{c}_{n'}^\dagger\}=0$; $\{\hat{c}_n^\dagger,\hat{c}_{n'}\}=\delta_{n,n'}$, where $\{\hat{a},\hat{b}\}=\hat{a}\,\hat{b}+\hat{b}\,\hat{a}$. We have to evaluate terms of the form $\overline{\hat{c}_{n_1}^\dagger\hat{c}_{n_2}\hat{c}_{n_3}^\dagger\hat{c}_{n_4}}=\sum_r P_r\langle r|\hat{c}_{n_1}^\dagger\hat{c}_{n_2}\hat{c}_{n_3}^\dagger\hat{c}_{n_4}|r\rangle$, where r denotes a particular Fock state and P_r is the canonical probability that the system is in state r. Given that $\langle r|c_n|r\rangle=\langle r|c_n^\dagger|r\rangle=0$ and $\langle r|c_n^\dagger c_n|r\rangle=f_{\mathrm{FD}}(n)$, where $f_{\mathrm{FD}}(n)$ is the Fermi-Dirac distribution evaluated at energy $\hbar\omega_n$, the correlation is of the form

$$\begin{aligned}\overline{\hat{c}_{n_1}^\dagger\hat{c}_{n_2}\hat{c}_{n_3}^\dagger\hat{c}_{n_4}}=&\,\delta_{n_1,n_2}\delta_{n_3,n_4}(1-\delta_{n_1,n_3})C_1\\&+\delta_{n_1,n_4}\delta_{n_3,n_2}(1-\delta_{n_1,n_2})C_2\\&+\delta_{n_1,n_2}\delta_{n_1,n_3}\delta_{n_1,n_4}C_3.\end{aligned} \tag{B5}$$

Here, $C_1=f_{\mathrm{FD}}(n_1)f_{\mathrm{FD}}(n_3)$, $C_2=f_{\mathrm{FD}}(n_1)[1-f_{\mathrm{FD}}(n_2)]$, and $C_3=f_{\mathrm{FD}}(n_1)$. We note that the terms C_1 and C_3 contribute only to a zero frequency [because Eq. (B4) implies the condition $n_1\neq n_2$ and $n_3\neq n_4$ to have nonzero ω and ω' frequencies in the $\overline{\hat{j}_p(\mathbf{r},\omega)\hat{j}_q(\mathbf{r}',\omega')}$ average] so that only the contribution C_2 proportional to $f_{\mathrm{FD}}(n_1)[1-f_{\mathrm{FD}}(n_2)]$ yields a nonzero contribution to the radiated field. Furthermore, considering that the Fermi level on the left side is larger than the Fermi level on the right side, Eq. (B4) selects the terms where $n_1=(\mathbf{k}_L,L)$ and $n_2=(\mathbf{k}_R,R)$ so that $\omega=\omega_{n_1}-\omega_{n_2}=\omega_L^{\mathrm{el}}-\omega_R^{\mathrm{el}}>0$. Hence, the statistical average is nonzero only if there is an electron-hole pair with an electron in a state L and a hole in a state R. Finally, we obtain the cross-spectral density of the current density:

$$\mathcal{S}_{j_pj_q}(\mathbf{r},\mathbf{r}',\omega)=\sum_{\mathbf{k}_L}\sum_{\mathbf{k}_R}j_{p,L\to R}(\mathbf{r})j_{q,L\to R}^*(\mathbf{r}')2\pi\delta(\omega-\omega_L^{\mathrm{el}}+\omega_R^{\mathrm{el}})f_{\mathrm{FD}}^L(\mathbf{k}_L)[1-f_{\mathrm{FD}}^R(\mathbf{k}_R)]. \tag{B6}$$

APPENDIX C: ORIGIN OF THE STRONG CURRENT CORRELATIONS AT OPPOSITE METALS

In Fig. 3(b) in the main text, we plot the cross-spectral density $\mathcal{S}_{j_zj_z}(z,z',\omega_{\mathbf{K}_\parallel}^{(\nu)})$ associated with the spatial correlations of the current density due to electrons tunneling through a narrow gap. This calculation is performed by introducing the wave functions of the QDS in the definition of $\mathcal{S}_{j_zj_z}(z,z',\omega_{\mathbf{K}_\parallel}^{(\nu)})$ from Eq. (27) and then solving the integral over $\mathbf{k}_L$ and $\mathbf{k}_R$ numerically. We discussed that the electronic currents are strongly correlated for $z=z'$ but also for points satisfying the condition $z'-L_{\mathrm{gap}}=-z\sqrt{(E_F^R+V_B/E_F^L)}$ at opposite metals. In this appendix, we analyze analytically the origin of the strong correlations.

The cross-spectral density $\mathcal{S}_{j_zj_z}(z,z',\omega_{\mathbf{K}_\parallel}^{(\nu)})$ depends on the expression of the current density $\mathbf{j}_{L\to R}$ associated to each particular transition [Eq. (19)]. We consider the QDS and, thus, use Eq. (2) for the wave function $\Psi_L(\mathbf{r})$ of a state in the left metal and Eq. (3) for the right metal. The final

state $\Psi_R(\mathbf{r})$ has a lower energy than the initial state $\Psi_L(\mathbf{r})$. By introducing their corresponding expression in Eq. (19), we obtain the following expressions for the z component of the current density. For $z \leq 0$:

$$\begin{aligned} j_{z,L\to R}(z) &= \frac{i\hbar e}{2m_{\text{eff}}}\frac{1}{L_z S} t_R^* e^{-ik_{zL}^- L_{\text{gap}}}\left[e^{i(k_{zL}^+ + k_{zL}^-)z} i(k_{zL}^+ - k_{zL}^-) - r_L e^{-i(k_{zL}^+ - k_{zL}^-)z} i(k_{zL}^+ + k_{zL}^-)\right] \\ &\approx \frac{\hbar e}{2m_{\text{eff}}}\frac{1}{L_z S} t_R^* e^{-ik_{zL}^- L_{\text{gap}}} r_L e^{-i(k_{zL}^+ - k_{zL}^-)z}(k_{zL}^+ + k_{zL}^-). \end{aligned} \tag{C1}$$

For $0 < z \leq L_{\text{gap}}$:

$$\begin{aligned} j_{z,L\to R}(z) = \frac{i\hbar e}{2m_{\text{eff}}}\frac{1}{L_z S}\Big[&(k_{z\text{gap}}^+ + k_{z\text{gap}}^-)\alpha_R^*\alpha_L e^{(k_{z\text{gap}}^+ - k_{z\text{gap}}^-)z} e^{k_{z\text{gap}}^- L_{\text{gap}}} - (k_{z\text{gap}}^+ + k_{z\text{gap}}^-)\beta_R^*\beta_L e^{-(k_{z\text{gap}}^+ - k_{z\text{gap}}^-)z} e^{-k_{z\text{gap}}^- L_{\text{gap}}} \\ &- (k_{z\text{gap}}^+ - k_{z\text{gap}}^-)\alpha_R^*\beta_L e^{-(k_{z\text{gap}}^+ + k_{z\text{gap}}^-)z} e^{k_{z\text{gap}}^- L_{\text{gap}}} + (k_{z\text{gap}}^+ - k_{z\text{gap}}^-)\beta_R^*\alpha_L e^{(k_{z\text{gap}}^+ + k_{z\text{gap}}^-)z} e^{-k_{z\text{gap}}^- L_{\text{gap}}}\Big]. \end{aligned} \tag{C2}$$

For $z > L_{\text{gap}}$:

$$\begin{aligned} j_{z,L\to R}(z) &= \frac{i\hbar e}{2m_{\text{eff}}}\frac{1}{L_z S} t_L\left[e^{i(k_{zR}^+ + k_{zR}^-)z} e^{-ik_{zR}^- L_{\text{gap}}} i(k_{zR}^+ - k_{zR}^-) + r_R^* e^{i(k_{zR}^+ - k_{zR}^-)z} e^{ik_{zR}^- L_{\text{gap}}} i(k_{zR}^+ + k_{zR}^-)\right] \\ &\approx -\frac{\hbar e}{2m_{\text{eff}}}\frac{1}{L_z S} t_L r_R^* e^{i(k_{zR}^+ - k_{zR}^-)z} e^{ik_{zR}^- L_{\text{gap}}}(k_{zR}^+ + k_{zR}^-). \end{aligned} \tag{C3}$$

In these expressions, the superscripts $+$ and $-$ in k_{zL}, k_{zR}, and $k_{z\text{gap}}$ mean that these wave vectors are associated to the state of the initial state (at higher energy) or to the final state (at lower energy), respectively. Furthermore, in Eqs. (C1) and (C3), we assume that $k_{zL(R)}^+ + k_{zL(R)}^- \gg |k_{zL(R)}^+ - k_{zL(R)}^-|$.

To calculate the cross-spectral density $\mathcal{S}_{j_z j_z}(z, z', \omega_{\mathbf{K}_\parallel}^{(\nu)})$, we need to integrate the product of the current density of Eqs. (C1)–(C3) at two points z and z', over all initial (with wave vector $\mathbf{k}_L$) and final (with wave vector $\mathbf{k}_R$) states, as indicated by Eq. (27). Crucially, to obtain strong correlations at different points z and z', the relation between the phases of the current density $j_{z,L\to R}(z)$ at these two points must be equal for all transitions between $\Psi_L(z)$ and $\Psi_R(z)$ states, so that contributions with different phases do not cancel out with each other. For each transition, we obtain the following phase relations of the current density over space [75]:

$$\arg\{j_{z,L\to R}(z)\} \approx \begin{cases} \arg(t_R^*) + \arg(r_L) - k_{zL}^- L_{\text{gap}} - (k_{zL}^+ - k_{zL}^-)z & z \leq 0, \\ \text{const} & 0 < z \leq L_{\text{gap}}, \\ \pi + \arg(t_L) + \arg(r_R^*) + k_{zR}^- L_{\text{gap}} + (k_{zR}^+ - k_{zR}^-)z & L_{\text{gap}} < z, \end{cases} \tag{C4}$$

where arg indicates the argument (or phase) of a complex number. In the length scales of usual insulator gaps, the argument of Eq. (C2) changes very slightly, and, thus, we have considered in the derivation of Eq. (C4) that it is constant over the gap region. Importantly, the phase of the current density changes linearly with z in the metals, due to the exponential terms $e^{-i(k_{zL}^+ - k_{zL}^-)z}$ and $e^{i(k_{zR}^+ - k_{zR}^-)z}$ in Eqs. (C1) and (C3), respectively. Furthermore, the continuity of the wave functions $\Psi_L(z)$ and $\Psi_R(z)$ implies that $j_{z,L\to R}(z)$ must be also continuous. By defining ζ as the argument of the current density in the two metal-insulator boundaries at the gap, we rewrite Eq. (C4) as

$$\begin{aligned} &\arg\{j_{z,L\to R}(z)\} \\ &\approx \begin{cases} \zeta - (k_{zL}^+ - k_{zL}^-)z & z \leq 0, \\ \zeta & 0 < z \leq L_{\text{gap}}, \\ \zeta + (k_{zR}^+ - k_{zR}^-)(z - L_{\text{gap}}) & L_{\text{gap}} < z. \end{cases} \end{aligned} \tag{C5}$$

An important consequence of Eq. (C5) is that the values of $\arg\{j_{z,L\to R}(z)\}$ are related in the two metals. Indeed, at two points z and z' satisfying the condition

$$z = -\frac{k_{zR}^+ - k_{zR}^-}{k_{zL}^+ - k_{zL}^-}(z' - L_{\text{gap}}), \tag{C6}$$

the current density has the same argument. Considering a transition from an initial state of energy $\hbar\omega^{\mathrm{el}}$ and parallel wave vector $\mathbf{k}_{\parallel}^{+}$ to a final state of corresponding values $\hbar\omega^{\mathrm{el}} - \hbar\omega_{\mathbf{K}_{\parallel}}^{(\nu)}$ and $\mathbf{k}_{\parallel}^{-}$, the denominator of Eq. (C6) is evaluated as

$$\begin{aligned} k_{zL}^{+} - k_{zL}^{-} &= \sqrt{\frac{2m_{\mathrm{eff}}(\hbar\omega^{\mathrm{el}} + E_F^L)}{\hbar^2} - |\mathbf{k}_{\parallel}^{+}|^2} - \sqrt{\frac{2m_{\mathrm{eff}}(\hbar\omega^{\mathrm{el}} - \hbar\omega_{\mathbf{K}_{\parallel}}^{(\nu)} + E_F^L)}{\hbar^2} - |\mathbf{k}_{\parallel}^{-}|^2} \\ &= \sqrt{\frac{2m_{\mathrm{eff}}E_F^L}{\hbar^2}}\left(\sqrt{1 + \frac{\hbar\omega^{\mathrm{el}} - \frac{\hbar|\mathbf{k}_{\parallel}^{+}|^2}{2m_{\mathrm{eff}}}}{E_F^L}} - \sqrt{1 + \frac{\hbar\omega^{\mathrm{el}} - \hbar\omega_{\mathbf{K}_{\parallel}}^{(\nu)} - \frac{\hbar|\mathbf{k}_{\parallel}^{-}|^2}{2m_{\mathrm{eff}}}}{E_F^L}}\right) \\ &\approx \sqrt{\frac{2m_{\mathrm{eff}}E_F^L}{\hbar^2}}\frac{\hbar\omega_{\mathbf{K}_{\parallel}}^{(\nu)} + \frac{\hbar(|\mathbf{k}_{\parallel}^{-}|^2 - |\mathbf{k}_{\parallel}^{+}|^2)}{2m_{\mathrm{eff}}}}{E_F^L}, \end{aligned} \tag{C7}$$

where in the last step we make a first-order Taylor expansion under the assumption that the electronic energies $\hbar\omega^{\mathrm{el}} - (\hbar|\mathbf{k}_{\parallel}^{+}|^2/2m_{\mathrm{eff}})$ are considerably smaller than the Fermi energy E_F^L in the transitions considered. Following the same calculation for $k_{zR}^{+} - k_{zR}^{-}$, we obtain the same expression of Eq. (C7) with the substitution $E_F^L \to E_F^R + eV_B$. Accordingly, we can evaluate the fraction in Eq. (C6), which leads to the condition

$$z = -\sqrt{\frac{E_F^L}{E_F^R + eV_B}}(z' - L_{\mathrm{gap}}). \tag{C8}$$

Notably, under the approximations considered in this appendix, Eq. (C8) does not depend on the transition frequency $\omega_{\mathbf{K}_{\parallel}}^{(\nu)}$ and on the parallel wave vectors $\mathbf{k}_{\parallel}^{+}$ and $\mathbf{k}_{\parallel}^{-}$ of the initial and final quantum states. This means that, for all possible transitions from a state $\Psi_L(z)$ to a state $\Psi_R(z)$, the current density has a similar argument at points z and z' that satisfy Eq. (C8). Therefore, all transitions act constructively in the integral of Eq. (27) at these two points, leading to a strong peak in the correlations of the current density.

Last, we note that in Eq. (C6) we assume that the positions satisfy the conditions $z < 0$ and $z' > 0$. Thus, Eq. (C8) is valid only in this region, as can be observed in Fig. 3(b). For the $z > 0$ and $z' < 0$ regions, on the other hand, we can follow a similar argument starting from Eq. (C5), and we obtain the second expression shown in Fig. 3(b), i.e.,

$$z' = -(z - L_{\mathrm{gap}})\sqrt{\frac{E_F^L}{E_F^R + eV_B}}. \tag{C9}$$

APPENDIX D: EXCITATION RATE OF THE INTERMEDIATE VELOCITY MODE

Figure 7 shows the emission rate of the intermediate velocity SPP mode under varying bias potentials ($V_B = 0.6$, 1.2, 1.8, and 2.4 V), for a junction with thicknesses $L_{\mathrm{Al}} = 10$ nm, $L_{\mathrm{Au}} = 20$ nm, and $L_{\mathrm{gap}} = 3$ nm. The goal is to complete the results obtained for the fast and slow SPP modes in the main text. The calculation with the QDS ($P_{\mathrm{QDS}}^{(m)}$, results shown by solid lines) in the complete device reveals some oscillations as a function of the SPP energy. These oscillations are due to constructive or destructive interferences between the metal and gap contributions, as we have already discussed for the slow and fast modes. Curiously, the QDS predicts an overall smaller power transferred to the intermediate velocity mode than

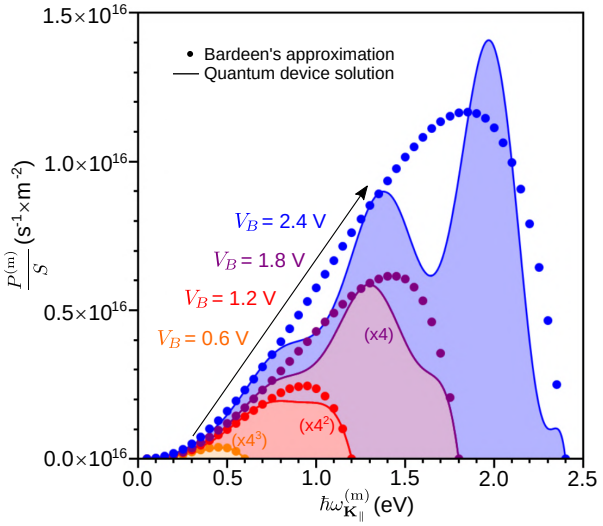

FIG. 7. Spectral nonradiative power $P^{(m)}(\hbar\omega_{\mathbf{K}_{\parallel}}^{(m)})$ per surface area S of the junction caused by the excitation of the intermediate velocity SPP mode, with thicknesses $L_{\mathrm{Al}} =$ 10 nm, $L_{\mathrm{Au}} = 20$ nm, and $L_{\mathrm{gap}} = 3$ nm [sketch in Fig. 4(a)] and bias potentials $V_B = 0.6$ (orange), 1.2 (red), 1.8 (purple), and 2.4 V (blue). Dots indicate the calculation using Bardeen's approximation, and the solid lines refer to the full calculation within the QDS.

Bardeen's approximation ($P_{\mathrm{BA}}^{(m)}$, results shown by dots). This is the opposite behavior found for the slow and fast modes (Figs. 5 and 6), where the QDS generally gives a larger contribution than Bardeen's approximation. Furthermore, the intermediate velocity mode excitation rate is much smaller than the slow mode excitation rate. In addition, it can be coupled only radiatively by roughness as opposed to the fast mode. Therefore, the intermediate velocity mode is likely negligible regarding light emission by a planar junction, so that the contribution of the metallic electrodes can be expected to increase the total emission rate once all SPP modes are considered.

APPENDIX E: SURFACE CONTRIBUTION IN THE EXCITATION OF THE FAST MODE

In the QDS approach to calculate the SPP excitation rate in planar junctions, we need to calculate the integral of Eq. (18). The regions of this integral are delimited by the volumes of the insulator gap (V_{gap}) and of the metals (V_{met}). In most of the calculations of the main text [except in Fig. 6(e)], we consider a local approach of electromagnetism to calculate the electric field of the SPPs. This perspective implies that in the metallic regions (inside the volume V_{met}) the system has a Drude permittivity and that outside this region the permittivity is constant and positive. Accordingly, the electromagnetic fields drop strongly exactly at the interface between these two regions, so that the electronic current interacts with the electric field in the gap and the weaker electric field inside the metallic regions, while it does not interact with the strong electric fields in vacuum and in the substrate (see Fig. 4).

However, according to the nonlocal theory of optical response of metals, the tunneling electrons could interact with the strong electric fields close to the metal-vacuum boundary. To visualize this, we show in Fig. 8(a) a sketch of the electronic density distribution close to this boundary. As an example, we consider a simple jellium model for the metal, where the ions create an uniform density of positive charge (indicated by the blue area). This uniform background leads to the calculation of the electronic ground state density $\rho_0(z)$ that decays gradually at the boundary between the jellium charge density and vacuum (gray region). $\rho_0(z)$ indicates the region where the wave functions of the tunneling electrons are well defined. Regarding the fields at optical frequency, the discontinuity of the normal fields takes place across a finite but narrow region where an oscillating charge density is induced. This region of strong induced charges is called the centroid of charge. The position of the centroid of charge does not coincide with the interface between the jellium and vacuum. It can be either inside [case 1 indicated by $\rho_{\mathrm{ind1}}(z)$, brown curve] or outside [case 2 indicated by $\rho_{\mathrm{ind2}}(z)$, green curve] the jellium edge,

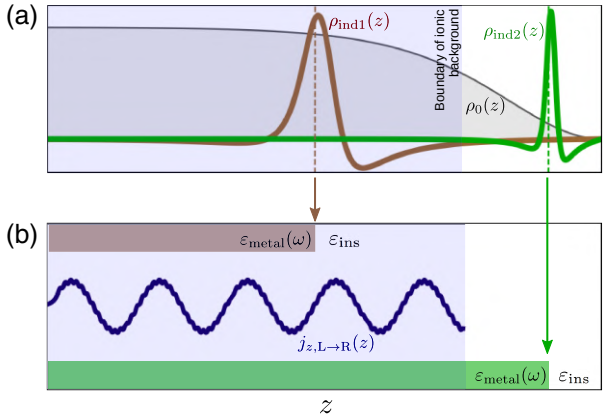

FIG. 8. (a) Schematics of the ground state density $\rho_0(z)$ (gray) and the induced charge densities for two different possible metals [$\rho_{\mathrm{ind1}}(z)$ in brown and $\rho_{\mathrm{ind2}}(z)$ in green]. The dashed lines indicate position of the centroid of the induced charge density of the same color. The blue area indicates the steplike positive charge background that is considered in the jellium model of metals. (b) Schematics of the distribution of the electronic current density $j_{z,L\to R}(z)$ considered in the calculations (solid blue line), up to the boundary of the ionic positive background. We also show the spatial distribution of the permittivities ($\varepsilon_{\mathrm{metal}}$ for the metal and $\varepsilon_{\mathrm{ins}}$ for the insulator) considered in our nonlocal approach for each $\rho_{\mathrm{ind}}(z)$ distribution shown in (a).

depending on the metal (inside for noble metals including d-band excitations, such as Au or Ag, and outside for s-like metals such as Na or Al). The positions of the centroid of the charge density are highlighted by vertical dashed lines in the figure. From the optical point of view, a simple model to obtain the optical response amounts to shifting the position of the metal interface to the position given by the charge centroid. This shift is given by the so-called Feibelman parameters [69–72]. The key point regarding the interaction between tunneling electrons and optical modes is that in case 1 (brown curve) the tunneling electrons (present in the jellium) can interact with an electric field lying outside the centroid of charge and, therefore, much larger.

To account for this interaction, we use a very simple approach. We consider that the electronic current is defined in the region delimited by the metallic volume V_{met} of the ionic background in the jellium model, as shown by the blue curve (current density $j_{z,L\to R}$) and the blue area (ionic background) in Fig. 8(b). On the other hand, when calculating the electric fields of the SPPs, we assume that the boundary between the metallic permittivity $\varepsilon_{\mathrm{metal}}(\omega)$ and the vacuum permittivity is given by the centroid of the induced charge density (as shown by brown and green). Importantly, the boundary between permittivities generally does not coincide with the integration boundary of Eq. (18) given by V_{met} (blue area), and the difference z_0 between them lies within a few angstroms (in our calculations, V_{met} extends outside the boundary between permittivities).

In the calculations for Fig. 6(e), we perform the integral in Eq. (18) by increasing its boundary along the z axis with respect to the boundary between the metallic and vacuum permittivities, and we show that with an increase of just a few angstroms the calculated excitation rate of the fast mode increases significantly.

[1] J. Lambe and S. L. McCarthy, *Light emission from inelastic electron tunneling*, Phys. Rev. Lett. **37**, 923 (1976).

[2] S. McCarthy and J. Lambe, *Enhancement of light emission from metal-insulator-metal tunnel junctions*, Appl. Phys. Lett. **30**, 427 (1977).

[3] B. Laks and D. L. Mills, *Light emission from tunnel junctions: The role of the fast surface polariton*, Phys. Rev. B **22**, 5723 (1980).

[4] J. Kirtley, T. N. Theis, and J. C. Tsang, *Light emission from tunnel junctions on gratings*, Phys. Rev. B **24**, 5650 (1981).

[5] P. Dawson, D. G. Walmsley, H. A. Quinn, and A. J. L. Ferguson, *Observation and explanation of light-emission spectra from statistically rough* Cu, Ag, and Au *tunnel junctions*, Phys. Rev. B **30**, 3164 (1984).

[6] S. Ushioda, J. E. Rutledge, and R. M. Pierce, *Theory of prism-coupled light emission from tunnel junctions*, Phys. Rev. B **34**, 6804 (1986).

[7] J. Gimzewski, B. Reihl, J. Coombs, and R. Schlittler, *Photon emission with the scanning tunneling microscope*, Z. Phys. B **72**, 497 (1988).

[8] J. K. Gimzewski, J. K. Sass, R. R. Schlitter, and J. Schott, *Enhanced photon emission in scanning tunnelling microscopy*, Europhys. Lett. **8**, 435 (1989).

[9] R. Berndt, R. Gaisch, J. K. Gimzewski, B. Reihl, R. R. Schlittler, W. D. Schneider, and M. Tschudy, *Photon emission at molecular resolution induced by a scanning tunneling microscope*, Science **262**, 1425 (1993).

[10] J. Aizpurua, G. Hoffmann, S. P. Apell, and R. Berndt, *Electromagnetic coupling on an atomic scale*, Phys. Rev. Lett. **89**, 156803 (2002).

[11] C. Chen, C. A. Bobisch, and W. Ho, *Visualization of Fermi's golden rule through imaging of light emission from atomic silver chains*, Science **325**, 981 (2009).

[12] Y. Zhang *et al.*, *Visualizing coherent intermolecular dipole-dipole coupling in real space*, Nature (London) **531**, 623 (2016).

[13] A. Rosławska, T. Neuman, B. Doppagne, A. G. Borisov, M. Romeo, F. Scheurer, J. Aizpurua, and G. Schull, *Mapping Lamb, Stark, and Purcell effects at a chromophore-picocavity junction with hyper-resolved fluorescence microscopy*, Phys. Rev. X **12**, 011012 (2022).

[14] P. Bharadwaj, A. Bouhelier, and L. Novotny, *Electrical excitation of surface plasmons*, Phys. Rev. Lett. **106**, 226802 (2011).

[15] T. Wang, E. Boer-Duchemin, Y. Zhang, G. Comtet, and G. Dujardin, *Excitation of propagating surface plasmons with a scanning tunnelling microscope*, Nanotechnology **22**, 175201 (2011).

[16] M. Parzefall, P. Bharadwaj, A. Jain, T. Taniguchi, K. Watanabe, and L. Novotny, *Antenna-coupled photon emission from hexagonal boron nitride tunnel junctions*, Nat. Nanotechnol. **10**, 1058 (2015).

[17] W. Du, T. Wang, H.-S. Chu, L. Wu, R. Liu, S. Sun, W. K. Phua, L. Wang, N. Tomczak, and C. A. Nijhuis, *On-chip molecular electronic plasmon sources based on self-assembled monolayer tunnel junctions*, Nat. Photonics **10**, 274 (2016).

[18] H. Qian, S.-W. Hsu, K. Gurunatha, C. T. Riley, J. Zhao, D. Lu, A. R. Tao, and Z. Liu, *Efficient light generation from enhanced inelastic electron tunnelling*, Nat. Photonics **12**, 485 (2018).

[19] C. Zhang, J.-P. Hugonin, A.-L. Coutrot, C. Sauvan, F. Marquier, and J.-J. Greffet, *Antenna surface plasmon emission by inelastic tunneling*, Nat. Commun. **10**, 4949 (2019).

[20] M. Parzefall, Á. Szabó, T. Taniguchi, K. Watanabe, M. Luisier, and L. Novotny, *Light from van der Waals quantum tunneling devices*, Nat. Commun. **10**, 292 (2019).

[21] Z. Wang, V. Kalathingal, T. X. Hoang, H.-S. Chu, and C. A. Nijhuis, *Optical anisotropy in van der Waals materials: Impact on direct excitation of plasmons and photons by quantum tunneling*, Light **10**, 230 (2021).

[22] J. H. Coombs, J. K. Gimzewski, B. Reihl, J. K. Sass, and R. R. Schlittler, *Photon emission experiments with the scanning tunnelling microscope*, J. Microscopy **152**, 325 (1988).

[23] A. Vilan, *Analyzing molecular current-voltage characteristics with the Simmons tunneling model: Scaling and linearization*, J. Phys. Chem. C **111**, 4431 (2007).

[24] S. Banerjee and P. Zhang, *A generalized self-consistent model for quantum tunneling current in dissimilar metal-insulator-metal junction*, AIP Adv. **9**, 085302 (2019).

[25] A. V. Uskov, J. B. Khurgin, I. E. Protsenko, I. V. Smetanin, and A. Bouhelier, *Excitation of plasmonic nanoantennas by nonresonant and resonant electron tunnelling*, Nanoscale **8**, 14573 (2016).

[26] H. Qian, S. Li, S.-W. Hsu, C.-F. Chen, F. Tian, A. R. Tao, and Z. Liu, *Highly-efficient electrically-driven localized surface plasmon source enabled by resonant inelastic electron tunneling*, Nat. Commun. **12**, 3111 (2021).

[27] M. H. Devoret, D. Esteve, H. Grabert, G. L. Ingold, H. Pothier, and C. Urbina, *Effect of the electromagnetic environment on the Coulomb blockade in ultrasmall tunnel junctions*, Phys. Rev. Lett. **64**, 1824 (1990).

[28] G.-L. Ingold and Y. V. Nazarov, *Charge Tunneling Rates in Ultrasmall Junctions* (Springer, Boston, 1992), pp. 21–107.

[29] M. Hofheinz, F. Portier, Q. Baudouin, P. Joyez, D. Vion, P. Bertet, P. Roche, and D. Esteve, *Bright side of the Coulomb blockade*, Phys. Rev. Lett. **106**, 217005 (2011).

[30] M. Hänisch and A. Otto, *Light emission from rough tunnel junctions in UHV*, J. Phys. Condens. Matter **6**, 9659 (1994).

[31] J. R. Kirtley, T. N. Theis, J. C. Tsang, and D. J. DiMaria, *Hot-electron picture of light emission from tunnel junctions*, Phys. Rev. B **27**, 4601 (1983).

[32] A. Mooradian, *Photoluminescence of metals*, Phys. Rev. Lett. **22**, 185 (1969).

[33] B. N. J. Persson and A. Baratoff, *Theory of photon emission in electron tunneling to metallic particles*, Phys. Rev. Lett. **68**, 3224 (1992).

[34] P. Johansson, R. Monreal, and P. Apell, *Theory for light emission from a scanning tunneling microscope*, Phys. Rev. B **42**, 9210 (1990).

[35] B. Laks and D. L. Mills, *Photon emission from slightly roughened tunnel junctions*, Phys. Rev. B **20**, 4962 (1979).

[36] J. Bardeen, *Tunnelling from a many-particle point of view*, Phys. Rev. Lett. **6**, 57 (1961).

[37] L. C. Davis, *Theory of surface-plasmon excitation in metal-insulator-metal tunnel junctions*, Phys. Rev. B **16**, 2482 (1977).

[38] M. Parzefall and L. Novotny, *Optical antennas driven by quantum tunneling: A key issues review*, Rep. Prog. Phys. **82**, 112401 (2019).

[39] D. Hone, B. Mühlschlegel, and D. J. Scalapino, *Theory of light emission from small particle tunnel junctions*, Appl. Phys. Lett. **33**, 203 (1978).

[40] C. B. Duke, *Tunneling in Solids* (Academic, New York, 1969), Vol. 10.

[41] A. D. Gottlieb and L. Wesoloski, *Bardeen's tunnelling theory as applied to scanning tunnelling microscopy: A technical guide to the traditional interpretation*, Nanotechnology **17**, R57 (2006).

[42] J. J. Sakurai, *Modern Quantum Mechanics* (Addison-Wesley, Reading, MA, 1994).

[43] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders, Philadelphia, 1976).

[44] M. Groner, J. Elam, F. Fabreguette, and S. George, *Electrical characterization of thin Al_2O_3 films grown by atomic layer deposition on silicon and various metal substrates*, Thin Solid Films **413**, 186 (2002).

[45] A. Downes, M. E. Taylor, and M. E. Welland, *Two-sphere model of photon emission from the scanning tunneling microscope*, Phys. Rev. B **57**, 6706 (1998).

[46] A. Archambault, F. Marquier, J.-J. Greffet, and C. Arnold, *Quantum theory of spontaneous and stimulated emission of surface plasmons*, Phys. Rev. B **82**, 035411 (2010).

[47] A. Archambault, *Optique des ondes de surface: Super-résolution et interaction matière-rayonnement*, theses, Université Paris Sud—Paris XI, 2011.

[48] P. Johansson and R. Monreal, *Theory for photon emission from a scanning tunneling microscope*, Z. Phys. B **84**, 269 (1991).

[49] J. Aizpurua, S. P. Apell, and R. Berndt, *Role of tip shape in light emission from the scanning tunneling microscope*, Phys. Rev. B **62**, 2065 (2000).

[50] K. Arya and R. Zeyher, *Light emission from tunnel junctions: The role of multiple scattering of surface polaritons*, Phys. Rev. B **28**, 4080 (1983).

[51] Y. Uehara, Y. Kimura, and S. U. Takeuchi, *Theory of visible light emission from scanning tunneling microscope*, Jpn. J. Appl. Phys. **31**, 2465 (1992).

[52] F. Bigourdan, J.-P. Hugonin, F. Marquier, C. Sauvan, and J.-J. Greffet, *Nanoantenna for electrical generation of surface plasmon polaritons*, Phys. Rev. Lett. **116**, 106803 (2016).

[53] P. Février and J. Gabelli, *Tunneling time probed by quantum shot noise*, Nat. Commun. **9**, 4940 (2018).

[54] J.-J. Greffet, M. Laroche, and F. Marquier, *Impedance of a nanoantenna and a single quantum emitter*, Phys. Rev. Lett. **105**, 117701 (2010).

[55] J. Lambe and R. Jaklevic, *Molecular vibration spectra by inelastic electron tunneling*, Phys. Rev. **165**, 821 (1968).

[56] L. Novotny and B. Hecht, *Principles of Nano-Optics* (Cambridge University Press, Cambridge, England, 2012).

[57] H. Benisty, J. Greffet, and P. Lalanne, *Introduction to Nanophotonics*, Oxford Graduate Texts (Oxford University, New York, 2022).

[58] T. Zhao *et al.*, *Plasmon dephasing in gold nanorods studied using single-nanoparticle interferometric nonlinear optical microscopy*, J. Phys. Chem. C **120**, 4071 (2016).

[59] A. Martin-Jimenez, A. I. Fernández-Domínguez, K. Lauwaet, D. Granados, R. Miranda, F. J. García-Vidal, and R. Otero, *Unveiling the radiative local density of optical states of a plasmonic nanocavity by STM*, Nat. Commun. **11**, 1021 (2020).

[60] C. C. Leon, A. Rosławska, A. Grewal, O. Gunnarsson, K. Kuhnke, and K. Kern, *Photon superbunching from a generic tunnel junction*, Sci. Adv. **5**, eaav4986 (2019).

[61] Y. Muniz, F. S. S. da Rosa, C. Farina, D. Szilard, and W. J. M. Kort-Kamp, *Quantum two-photon emission in a photonic cavity*, Phys. Rev. A **100**, 023818 (2019).

[62] M. Parzefall and L. Novotny, *Light at the end of the tunnel*, ACS Photonics **5**, 4195 (2018).

[63] M. A. Ordal, R. J. Bell, R. W. Alexander, L. L. Long, and M. R. Querry, *Optical properties of fourteen metals in the infrared and far infrared:* Al, Co, Cu, Au, Fe, Pb, Mo, Ni, Pd, Pt, Ag, Ti, *V*, and *W*, Appl. Opt. **24**, 4493 (1985).

[64] R. Esteban, A. Zugarramurdi, P. Zhang, P. Nordlander, F. J. García-Vidal, A. G. Borisov, and J. Aizpurua, *A classical treatment of optical tunneling in plasmonic gaps: Extending the quantum corrected model to practical situations*, Faraday Discuss. **178**, 151 (2015).

[65] D. Gall, *Electron mean free path in elemental metals*, J. Appl. Phys. **119**, 085101 (2016).

[66] E. N. Economou, *Surface plasmons in thin films*, Phys. Rev. **182**, 539 (1969).

[67] J. M. Pitarke, V. M. Silkin, E. V. Chulkov, and P. M. Echenique, *Theory of surface plasmons and surface-plasmon polaritons*, Rep. Prog. Phys. **70**, 1 (2006).

[68] G. Shalem, O. Erez-Cohen, D. Mahalu, and I. Bar-Joseph, *Light emission in metal-semiconductor tunnel junctions: Direct evidence for electron heating by plasmon decay*, Nano Lett. **21**, 1282 (2021).

[69] P. J. Feibelman, *Surface electromagnetic fields*, Prog. Surf. Sci. **12**, 287 (1982).

[70] T. V. Teperik, P. Nordlander, J. Aizpurua, and A. G. Borisov, *Robust subnanometric plasmon ruler by rescaling of the nonlocal optical response*, Phys. Rev. Lett. **110**, 263901 (2013).

[71] P. Gonçalves, T. Christensen, N. Rivera, A.-P. Jauho, N. A. Mortensen, and M. Soljačić, *Plasmon-emitter interactions at the nanoscale*, Nat. Commun. **11**, 366 (2020).

[72] N. A. Mortensen *et al.*, *Surface-response functions obtained from equilibrium electron-density profiles*, Nanophotonics **10**, 3647 (2021).

[73] A. Babaze, E. Ogando, P. Elli Stamatopoulou, C. Tserkezis, N. A. Mortensen, J. Aizpurua, A. G. Borisov, and R. Esteban, *Quantum surface effects in the electromagnetic coupling between a quantum emitter and a plasmonic nanoantenna: Time-dependent density functional theory vs. semiclassical Feibelman approach*, Opt. Express **30**, 21159 (2022).

[74] L. Giuliani and G. Vignale, *Quantum Theory of the Electron Liquid* (Cambridge University Press, Cambridge, England, 2005).

[75] In all equations involving the arg function in this appendix, the equalities are satisfied under mod(2π).