Plasmon-assisted resonant tunneling in graphene-based heterostructures

V. Enaldiev,[1] A. Bylinkin,[1] and D. Svintsov[1,2]
[1]*Laboratory of 2d Materials' Optoelectronics, Moscow Institute of Physics and Technology, Dolgoprudny 141700, Russia*
[2]*Laboratory of Sub-micron Devices, Institute of Physics and Technology RAS, Moscow 117218, Russia*
(Received 11 July 2017; published 28 September 2017)

We develop a theory of plasmon-assisted tunneling in graphene-insulator-graphene heterostructures and reveal the manifestations of such process in current-voltage curves, plasmon emission spectra, and junction electroluminescence. We present a unified framework for evaluation of tunneling due to carrier-carrier Coulomb scattering and due to emission of plasmons; the latter mechanism generally dominates the full inelastic current. Moreover, the plasmon-assisted current and plasmon emission rates possess resonant peaks at voltages providing equal energies, momenta, and group velocities of collective and single-particle interlayer excitations. This resonance is unique to the tunnel-coupled 2d systems of massless electrons and is deeply related to strong interactions between collinear carriers in graphene. The predicted effect can be used for design of efficient nanoscale voltage-tunable sources of photons and surface plasmons.

DOI: 10.1103/PhysRevB.96.125437

I. INTRODUCTION

Van der Waals heterostructures composed of graphene and related 2d materials [1] provide a new platform for both fundamental studies of solid state [2–4] and novel optoelectronic devices [5–8]. A basic building block of these structures is the graphene-insulator-graphene (GIG) sandwich which, despite its apparent simplicity, keeps on demonstrating intriguing properties. The two layers in a GIG structure act as gates for each other enabling the voltage control of electronic and optical [9,10] properties. As the layers are placed closer, the Coulomb interaction between remote electrons comes into play, enabling interlayer drag [11]. This effect still puzzles researchers with anomalies at the charge neutrality point [12–14] which still lack an established explanation. Further progress in formation of ultrathin high-quality barrier layers has enabled the observation of chiral electron resonant tunneling [15], selective valley injection [2], and electrical tuning of carrier thermalization [4] in GIG structures.

The experimental data on electron tunneling in GIG structures [16–18] were for a long time described by a single-electron picture [19–21], with a minor revision due to electron-phonon interaction [22,23]. However, recent experiments have revealed the gate-controlled electroluminescence of graphene tunnel junctions correlating with negative differential resistance in the current-voltage [$I(V)$] curves [24]. A common scenario of electroluminescence in tunnel junctions involves the excitation of plasmons upon tunneling followed by their radiative decay [25–27]. Thus, the graphene junction luminescence hints at the possibility of plasmon-assisted tunneling. Though the properties of plasmons in graphene double layers are well established [10,28–31], and various tunneling mechanisms have been addressed (elastic including the interaction corrections [32], phonon- [23] and photon-assisted [33]), no theory is available for plasmon-assisted tunneling.

We develop such a theory in the present paper. In contrast to a common model which treats tunneling due to zero-point and thermal fluctuations of electric field in plasmon modes [34–36], we start with tunneling due to carrier-carrier scattering (Fig. 1). The dynamic screening of the Coulomb interaction resonantly enhances the probability of scattering-assisted tunneling if the transferred energy and momentum coincide with those of surface plasmons. The plasmon-pole contribution to scattering-aided current coincides with that obtained using the "fluctuation" approach in the limit of weak electromagnetic dissipation. However, our approach allows a consistent treatment of plasmon damping and gives the full inelastic current as an added benefit.

In ordinary tunnel junctions, the onset of plasmon-assisted tunneling is marked by a cusp in the differential conductivity dI/dV at the offset $eV_{\rm th}$ equal to the plasmon energy $\hbar\omega_p \propto n^{1/2}$, where n is the carrier density [37]. One can expect a similar picture, but with $eV_{\rm th} \propto n^{1/4}$, for GIG junctions [22]. Here, we show that the situation is very different. First, the threshold voltages for plasmon-assisted tunneling in graphene depend on carrier densities even more weakly due to the softness of plasmon dispersion in two dimensions. Second, and most importantly, the plasmon-assisted tunneling results not only in cusps in the dI/dV curves but also in the full-scale resonances in the $I(V)$ dependence. These resonances correspond to the energy, momentum, and group velocity matching between plasmons and interlayer single-particle excitations. In some sense, this effect is similar to the formation of plasmarons in a single graphene layer due to the consonance between plasmon and electron motion [38]; for this reason, we call it *plasmaronic resonance*. The strength of resonance is sensitive to plasmon lifetime and interlayer twist, and in misoriented graphene layers the inelastic current above the threshold voltage represents an unstructured background. Conversely, in aligned layers the inelastic scattering-assisted resonant current can dominate over the elastic one, and the role of carrier scattering is not limited to the broadening of the main elastic resonance [32].

The resonance in plasmon-assisted current for aligned layers will be accompanied by a peak in plasmon emission rate. This resonant emission can set the principle of electrically switchable plasmon sources, which remain the only missing element in the chain "generation, guiding [3], detection [39]" required for plasmonic interconnects. As the plasmons can couple to photons, the junction electroluminescence in aligned

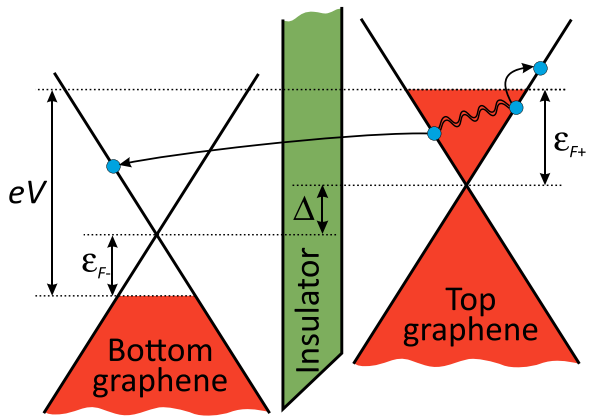

FIG. 1. Band diagram of a graphene-insulator-graphene tunnel junction and a schematic view of electron tunneling accompanied by electron-electron scattering. eV is the applied voltage, Δ is the interlayer band offset, and $\varepsilon_{F\pm}$ are the Fermi energies in the top (+) and bottom (−) layers.

layers will also be resonant. We estimate the radiative decay rate of surface plasmons in GIG junctions and show that the efficiency of plasmon-to-photon conversion can tend to unity if the junction is properly coupled to an antenna [27]. Therefore, the mechanism of tunneling under study looks prospective for the realization of ultracompact electrically tunable light sources.

The paper is organized as follows. In Sec. II A we develop a microscopic model of tunneling due to dynamically screened Coulomb interaction, and extract the contribution due to plasmon emission. In Sec. II B we show that the plasmon-assisted component of the current can be obtained by considering electron interactions with fluctuating electric fields. In Sec. III we discuss possible experimental manifestations of plasmon-assisted tunneling, including plasmaronic resonance (Sec. III A), the threshold structure of current-voltage curves (Sec. III B), and resonant electroluminescence (Sec. III C). In Sec. IV we discuss possible extensions of our model and draw the main conclusions. The details of the calculations are presented in the Appendices.

II. THEORY OF MANY-PARTICLE AND PLASMON-ASSISTED TUNNELING

Electron states in aligned tunnel-coupled graphene layers can be labeled by in-plane momentum $\mathbf{p}$, the band index $s = \pm 1$, and the index $l = \pm 1$ governing the vertical localization of the electron wave function [21]. The respective energies are $\epsilon^{ls}_{\mathbf{p}} = spv + l\sqrt{\Delta^2/4 + \Omega^2}$, where v is the Fermi velocity, Δ is the voltage-induced shift of bands in neighboring layers (band offset; see Fig. 1), and Ω is the tunnel splitting. For strong bias and/or weak tunneling, $\Delta \gg \Omega$, the state $l = +1$ can be regarded as belonging to the top layer and $l = -1$ – to the bottom one. The Fermi energies in the layers $\varepsilon_{F\pm}$ and the band offset Δ are related to the interlayer voltage via $eV = \Delta + \varepsilon_{F+} - \varepsilon_{F-}$ (Fig. 1). In principle, these quantities can be tuned independently with extra gates.

A. Scattering-assisted tunneling

The Coulomb interaction couples the states in neighboring layers and induces inelastic tunneling current. The current from the forward-biased (top) layer can be expressed through the golden rule transition probability W_{fi} between two-particle states $|+\mathbf{p}s,l\mathbf{p}_1s_1\rangle$ and $|-\mathbf{p}'s',l'\mathbf{p}'_1s'_1\rangle$ and occupation numbers $f^{ls}_{\mathbf{p}}$ of these states:

$$I_{t\to b} = eg^2 \sum_{\substack{\mathbf{pp_1q}\\ lss_1s'_1}} W_{fi} f^{+s}_{\mathbf{p}}\left[1 - f^{-s'}_{\mathbf{p}'}\right] f^{ls_1}_{\mathbf{p}_1}\left[1 - f^{ls'_1}_{\mathbf{p}'_1}\right]; \quad (1)$$

here $g = 4$ is the spin-valley degeneracy factor. The full current including the reverse component is $I = I_{t\to b}[1 - e^{-eV/kT}]$, where V is the interlayer voltage. A sequence of transformations common in the theory of Coulomb scattering phenomena [40,41] allows one to express the current in terms of the imaginary part of polarizability $\Pi''_{ll'}$ (we set $\hbar \equiv 1$):

$$I_{t\to b} = \frac{2e}{\pi}\int_{-\infty}^{+\infty} d\omega \sum_{\mathbf{q},l} \Pi''_{+-}(q,\omega)\Pi''_{ll}(q,\omega) \times |V_{+l,-l}|^2 N_{\omega - eV}[N_\omega + 1], \quad (2)$$

where $N_\omega = [e^{\omega/T} - 1]^{-1}$ is the Bose distribution and $\Pi_{ll'}$ is the intra- ($l = l'$) or interlayer ($l \neq l'$) polarizability in the random phase approximation [42,43]:

$$\Pi_{ll'} = \frac{g}{A}\sum_{\substack{\mathbf{pss'}\\ \mathbf{p'}=\mathbf{p}+\mathbf{q}}} \left|u^{ss'}_{\mathbf{pp'}}\right|^2 \frac{f^{ls}_{\mathbf{p}} - f^{l's'}_{\mathbf{p'}}}{\omega + i\delta - \epsilon^{ls}_{\mathbf{p}} + \epsilon^{l's'}_{\mathbf{p'}}}; \quad (3)$$

$|u^{ss'}_{\mathbf{pp'}}|^2 = (1 + ss'\cos\theta_{\mathbf{pp'}})/2$ is the overlap of chiral wave functions in graphene, and A is the sample area. The amplitude $V_{+l,-l}$ describes the tunneling of electrons from the top to bottom layer upon Coulomb interaction with a carrier in the lth layer. Its evaluation requires knowledge of electron wave functions in the vertical direction $\psi_l(z)$ and the full dynamic Coulomb interaction between two electrons $V_{\mathbf{q}\omega}(z,z')$:

$$V_{+l,-l} = \int_{-\infty}^{+\infty} |\psi_l(z')|^2 \psi_+(z) V_{\mathbf{q}\omega}(z,z')\psi_-(z)dzdz'. \quad (4)$$

The full Coulomb interaction is due to the bare charge and induced charges in both layers,

$$V_{\mathbf{q}\omega}(z,z') = V_0(\mathbf{q})e^{-q|z-z'|} + \sum_{l=\pm 1} V^{(S)}_{l\mathbf{q}\omega}(z_l - z), \quad (5)$$

where $V_0(\mathbf{q}) = 2\pi e^2/\kappa|\mathbf{q}|$, κ is the background dielectric constant, and the potential $V^{(S)}_{l\mathbf{q}\omega}(z_l - z)$ is due to screening by the dynamically induced density $\delta n_{l\mathbf{q}\omega}$ in the lth layer. The latter is related to the potential at the same layer via its polarizability $\Pi_{ll}(\mathbf{q},\omega)$.

Equations (2)–(5) are, in principle, sufficient to calculate the scattering-assisted current. Unluckily, the spatial structure of electron wave functions in Eq. (4) for van der Waals structures is not known. Considerable progress can be made if (1) one assumes the scattered electron to be well localized in its own layer and (2) considers the Coulomb potential in the dipole approximation. With these assumptions, all information about tunneling is neatly absorbed into the dipole matrix element $z_\pm$,

while the transition amplitude becomes

$$V_{+l,-l} = \frac{V_0(\mathbf{q})}{\varepsilon(\mathbf{q},\omega)}\frac{z_\pm}{d}(1-e^{-qd}) \times [1 - V_0(\mathbf{q})\Pi_{-l-l}(\mathbf{q},\omega)(1+e^{-qd})]; \quad (6)$$

here $\varepsilon(\mathbf{q},\omega)$ is the dynamic screening function of the double layer [28]:

$$\varepsilon(\mathbf{q},\omega) = [1 - V_0(q)\Pi_{++}(\mathbf{q},\omega)][1 - V_0(q)\Pi_{--}(\mathbf{q},\omega)] - V_0^2(q)e^{-2qd}\Pi_{++}(\mathbf{q},\omega)\Pi_{--}(\mathbf{q},\omega). \quad (7)$$

With Eq. (6) it becomes clear that the unscreened Coulomb interaction is inefficient for excitation of the tunneling transitions. This is elucidated by the appearance of prefactor $1-e^{-qd}$ which makes the full amplitude small unless the transferred momentum exceeds the inverse interlayer distance. Physically, this term captures the long-range nature of Coulomb interaction: despite that the potential created by a trial electron can be large, the *potential difference* between two layers is small unless the layers are distant. A similar cancellation of Coulomb scattering can lead to the narrowing of tunnel resonances in doped double quantum wells [44] and superlattices [45].

The situation changes radically due to the screening, which is seen from the resonant enhancement of matrix elements at $\varepsilon(\mathbf{q},\omega)\to 0$. This is nothing but tunneling accompanied by emission of surface plasmons. To extract the plasmonic contribution, one can expand the dielectric function in the vicinity of plasmon poles $\varepsilon(\mathbf{q},\omega) \approx [\partial\varepsilon'/\partial\omega](\omega-\omega_{\mathbf{q}}^p)+i\varepsilon''$ [here $p=+1$ (-1) corresponds to optical (acoustic) modes; for details see Appendix A]. Assuming the electromagnetic dissipation to be small, $|\varepsilon''/\varepsilon'| \ll 1$, we arrive at the expression for the *plasmon-assisted* component of the net tunneling current $I^{\rm pl}_{t\to b}$. This can be conveniently split into the emission ($I^{\rm pl,em}_{t\to b}$) and absorption ($I^{\rm pl,abs}_{t\to b}$) contributions

$$I^{\rm pl}_{t\to b} = I^{\rm pl,em}_{t\to b} + I^{\rm pl,abs}_{t\to b}, \quad (8)$$

$$I^{\rm pl,em}_{t\to b} = 2\pi e\sum_{\mathbf{q}p}\left|\frac{e\varphi^p_{\mathbf{q}\pm}}{2}\right|^2 \Pi''_{+-}\left(\mathbf{q},\omega^p_{\mathbf{q}}\right)\left[N_{\omega^p_{\mathbf{q}}}+1\right]N_{\omega^p_{\mathbf{q}}-eV}. \quad (9)$$

The quantity $e\varphi^p_{\mathbf{q}\pm}$ can be considered as a matrix element of electron-plasmon interaction:

$$\frac{\left(e\varphi^p_{\mathbf{q}}\right)^2_{(\pm)}}{2} = V_0(\mathbf{q})\left|\frac{z_\pm}{d}\right|^2(1-e^{-qd})^2 \times \frac{\left[1-V_0(\mathbf{q})\Pi'_{++}\left(\mathbf{q},\omega^p_q\right)(1+e^{-qd})\right]^2}{\frac{\partial\varepsilon'}{\partial\omega^p_q}\left|1-V_0(\mathbf{q})\Pi'_{++}\left(\mathbf{q},\omega^p_q\right)(1-e^{-2qd})\right|}. \quad (10)$$

Interestingly, this matrix element turns to zero for interactions with optical plasmon modes in equally doped layers. This fact stems from zero average field in the optical mode and vanishing tunnel coupling in the dipole approximation. A slight asymmetry in layer doping and/or unequal dielectric constants of the substrate and barrier layer leads to mixing of optical and acoustic modes. Therefore, both modes in asymmetric structures contribute to tunneling.

B. Tunneling due to electric field fluctuations

Equation (9) for tunneling current can be derived using the golden rule for electron-plasmon interactions. This approach provides us only with the part of the tunneling current arising from collective excitations, while the tunneling due to single-particle excitations is neglected. It also neglects the exchange effects in electron scattering, which are unimportant in graphene due to large spin-valley degeneracy. Finally, it implies that plasmons are well-defined; i.e., their damping is much less than the eigenfrequency. Despite all limitations, this approach is useful as the plasmon-aided current makes a considerable fraction of the full current, as we show below. Denoting the tunneling matrix element for electron-plasmon interaction as $e\varphi^p_{\mathbf{q}\pm}$, we can write the current from the top to bottom layer due to plasmon emission

$$I^{pl,em}_{t\to b} = 2\pi e\sum_{\mathbf{pq}ss'p} f^{+s}_{\mathbf{p}}\left(1-f^{-s'}_{\mathbf{p}-\mathbf{q}}\right)\left|e\varphi^p_{\mathbf{q}\pm}\right|^2\left|u^{ss'}_{\mathbf{pp}-\mathbf{q}}\right|^2 \times\left[N_{\omega^p_q}+1\right]\delta\left(\varepsilon^{+s}_{\mathbf{p}}-\varepsilon^{-s'}_{\mathbf{p}-\mathbf{q}}-\omega_q\right) \quad (11)$$

and plasmon absorption

$$I^{pl,\rm abs}_{t\to b} = 2\pi e\sum_{\mathbf{pq}ss'p} f^{+s}_{\mathbf{p}}\left(1-f^{-s'}_{\mathbf{p}+\mathbf{q}}\right)\left|e\varphi^p_{\mathbf{q}\pm}\right|^2\left|u^{ss'}_{\mathbf{pp}-\mathbf{q}}\right|^2 \times N_{\omega^p_q}\delta\left(\varepsilon^{+s}_{\mathbf{p}}-\varepsilon^{-s'}_{\mathbf{p}+\mathbf{q}}+\omega_q\right). \quad (12)$$

The speed of plasmons in 2d systems is generally well below the speed of light, which allows one to treat the plasmon field as a potential one. In this approximation, the tunneling matrix element takes on the form

$$e\varphi^p_{\mathbf{q}\pm} = \int_{-\infty}^{+\infty} dz\,\psi^*_+(z)e\varphi^p_{\mathbf{q}}(z)\psi_-(z), \quad (13)$$

where $\varphi^p_{\mathbf{q}}(z)$ describes the distribution of plasmon potential in the direction normal to the layers (see Appendix B). The amplitude of electric field fluctuations $e\varphi^p_{\mathbf{q}}(z)$ in the plasmonic modes is established with a second-quantization procedure, where the classical mode energy w is equated to $\omega^p_{\mathbf{q}}$ [34]. The classical energy of electromagnetic field is given by the Brillouin formula

$$w = \int d^3r\frac{\kappa\mathbf{E}\mathbf{E}^*}{16\pi} - \frac{A}{4}\sum_{l=\pm}\left.\frac{\partial\sigma''_l}{\partial\omega}\right|_{\omega^p_q}\left.\mathbf{E}_{||}\mathbf{E}^*_{||}\right|_{z=ld/2}, \quad (14)$$

where $\mathbf{E}=(\mathbf{E}_{||},E_z)=-(i\mathbf{q},\partial_z)\varphi^p_{\mathbf{q}}(z)$ is the electric field, A is the sample area, and σ_l is the conductivity of the lth layer. A set of lengthy transformations (Appendix B) leads us to the expression for the tunneling current in the form (8) with the matrix element (10).

The coincidence between the two approaches to plasmon-assisted tunneling is not accidental. It was shown already for classical plasmas that electron-electron collisions can be treated as electron scattering by longitudinal field fluctuations, their magnitude dictated by the fluctuation-dissipation theorem [46]. In nonequilibrium systems, the plasmonic occupation numbers in Eqs. (11) and (12) can be determined from the rate equations for plasmons.

III. MANIFESTATIONS OF PLASMON-ASSISTED TUNNELING

With the general formalism of the calculations developed, we start discussing the possible experimental manifestations of plasmon-assisted tunneling. We consider three such effects: (1) resonant enhancement of the tunnel current due to group velocity coincidence between plasmons and interlayer single-particle excitations, which we call *plasmaronic resonance*, (2) fine structure of the low-temperature $I(V)$ curves due to onset of the plasmon emission, and (3) electroluminescence of graphene tunnel junctions.

A. Plasmaronic resonance in tunnel current

To analyze the resonant structure of the tunnel current, Eq. (2), we note that the inter- [31] and intralayer [42,43] polarizabilities possess square-root singularities at the threshold of inter- and intralayer single-particle excitations, respectively. Particularly, the imaginary part of interlayer polarizability can be presented as

$$\Pi''_{+-}(\mathbf{q},\omega) = \frac{\widetilde{\Pi}''_{+-}(\mathbf{q},\omega)}{\sqrt{\pm[q^2v^2-(\omega-\Delta)^2]}}, \tag{15}$$

where $\widetilde{\Pi}$ is the smooth part, and the plus and minus signs should be used for intra- and interband transitions, respectively [47] (for details see Appendix C). This square-root singularity can be explained as resulting from prolonged interlayer interaction between electron and hole with collinear momenta and, hence, equal velocities. Similar singularities exist in the polarizability of a single graphene layer [42] $\Pi''_{ll}(\mathbf{q},\omega) \propto |q^2v^2-\omega^2|^{-1/2}$. The most pronounced effect of "collinear singularities" is that the plasmon *phase* velocity always exceeds the Fermi velocity, which preserves graphene plasmons from Landau damping [48,49].

The *group* velocity of plasmons can be, however, equal to or below the Fermi velocity. When the line of interlayer tunneling singularities $\omega = \Delta + qv$ approaches the tangent with plasmon dispersion, as shown in Fig. 2(b), the plasmon-assisted tunneling current is resonantly enhanced. The resonant interlayer band offset Δ^* is determined from

$$\Delta^* + qv = \omega^p_q, \quad \partial\omega^p_q/\partial q = v. \tag{16}$$

In the vicinity of resonance the plasmon-assisted contribution grows as

$$I^{\mathrm{pl,em}}_{t\to b}(\Delta) \approx I_0 \ln\left|\frac{q^2\partial^2\omega_q/\partial q^2}{2(\Delta-\Delta^*)}\right|\Bigg|_{q=q^*}, \tag{17}$$

where the large logarithm is evaluated at $q = q^*$ which is the momentum of plasmons in resonance with interlayer excitations. The characteristic current in Eq. (17) is

$$I_0 = \left|\frac{e\varphi^p_{\mathbf{q}\pm}}{2}\right|^2 \left.\frac{eq^3(N_\omega+1)N_{\omega-eV}\widetilde{\Pi}_\pm(q,\omega)}{2\pi\sqrt{qv\partial^2\omega_q/\partial q^2}}\right|_{\substack{q=q^*\\ \omega=\omega_{q^*}}}. \tag{18}$$

Both large logarithm and prefactor are growing functions of the plasmon dispersion curvature $\partial^2\omega_q/\partial q^2$. Indeed, when the curvature is low a wide range of plasmon wave vectors satisfies the resonant condition with single-particle excitations.

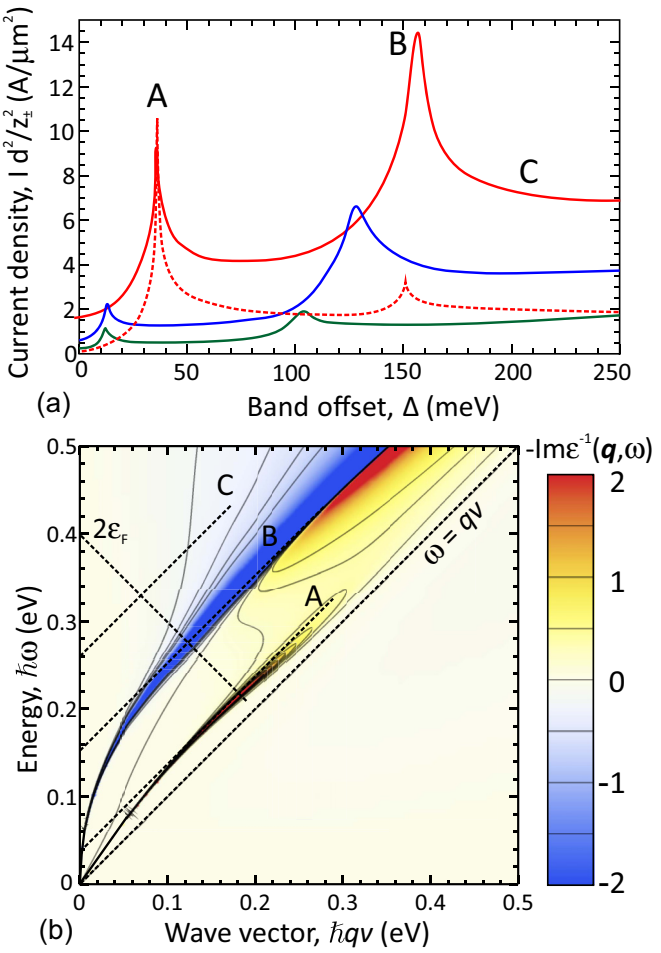

FIG. 2. (a) Calculated inelastic tunnel current (normalized by $z^2_\pm/d^2$) vs band offset Δ at fixed Fermi energies in graphene layers (red solid $\varepsilon_{F+} = 0.6$ eV, $\varepsilon_{F-} = -0.2$ eV; blue solid $\varepsilon_{F+} = 0.5$ eV, $\varepsilon_{F-} = -0.1$ eV; green solid $\varepsilon_{F+} = 0.4$ eV, $\varepsilon_{F-} = 0.1$ eV). Red dashed curve represents the plasmon-assisted current calculated with Eq. (8). Interlayer distance $d = 38$ Å, $\kappa = 5$, temperature $T = 300$ K. Peaks A and B correspond to plasmaronic resonances due to acoustic and optical modes, respectively. (b) Loss function $-\mathrm{Im}[\varepsilon^{-1}(\mathbf{q},\omega)]$ of the double-layer structure for the same parameters as for the red curve. Resonant peaks in the current correspond to the tangent of the interlayer excitations' dispersion $\omega = \Delta + qv$ (dashed line) and dispersion of surface plasmons (bright peaks in the loss function). The dot-dashed line is the boundary of interband absorption $\omega = 2\min\{\varepsilon_{F+},\varepsilon_{F-}\} - qv$.

The logarithmic growth of the current at the resonance is limited by plasmon damping. Using the many-particle formalism, Eq. (2), it is possible to show that the damping-limited resonant value of the current is approximately

$$I^{\mathrm{pl,em}}_{t\to b}(\Delta^*) \approx I_0 \ln\left|\frac{q^2\partial^2\omega_q/\partial q^2}{\varepsilon''/[\partial\varepsilon'/\partial\omega]}\right|\Bigg|_{q=q^*}; \tag{19}$$

the quantity in the denominator of the large logarithm is nothing but the plasmon decay rate.

The strength and width of plasmaronic resonances are not directly affected by temperature, contrary to the case of fine structures in $I(V)$ curves due to phonon-assisted tunneling

that require sharp Fermi edges [22,50]. However, the plasmon decay rate limiting the resonance strength can increase at elevated temperature. If plasmon damping is governed by the interband absorption and plasmon energy is far below the Fermi energy, then $\varepsilon''(q,\omega) \propto e^{(\omega+qv-2\varepsilon_F)/2T}$, and the maximum current is inversely proportional to the temperature.

A direct numerical evaluation of tunnel current with Eq. (2), Fig. 2, shows that the plasmaronic resonances can persist up to the room temperature. The two peaks in Fig. 2 correspond to the plasmaronic resonances on acoustic and optical modes. The dashed line shows the inelastic current in the plasmon-pole approximation, Eqs. (8) and (9). The latter correctly captures the peak positions and the magnitude of the acoustic peak but underestimates the net current due to the neglect of nonresonant transitions with $\omega \neq \omega_q$.

The plasmon-assisted current is proportional to a small parameter $z_\pm^2/d^2$ that contains all information about barrier material properties. It is possible to show (Appendix E) that $z_\pm/d \approx \Omega/\Delta \approx (2U_0/\Delta)e^{-\varkappa d}$, where U_0 is the band offset between graphene and barrier material, and $\varkappa = \sqrt{2m^*U_0}$ is the inverse decay length of the electron wave function. The same prefactor enters the expression for elastic current away from the resonance, $I_{el} = (2\pi e\gamma)(\Omega/\Delta)^2(n_+ + p_-)$, where γ is the resonance broadening factor, and n_+ and p_- are the electron and hole densities in the respective layers. Therefore, the ratio of plasmon-assisted and elastic currents is roughly independent of barrier layer parameters. In Fig. 7 we compare these currents for $d = 3.8$ nm of boron nitride and see that the resonant contribution due to acoustic plasmon emission exceeds the tail of the elastic current by nearly a factor of 3.

The inelastic tunneling resonances are highly sensitive to the rotational misalignment of graphene layers [17]. With the neglect of emerging weak tunneling between dissimilar sublattices, the general expression for inelastic current (2) still holds, but the interlayer polarizability is now angle-dependent. Denoting the wave vectors connecting the K points in the neighboring layers as $\mathbf{Q}_i$ $(i = 1\ldots3)$, we can write the polarizability in the presence of twist $\Pi_{+-}^{(T)}(\mathbf{q},\omega) = \frac{1}{3}\sum_{i=1\ldots3}\Pi_{+-}(\mathbf{q}+\mathbf{Q}_i,\omega)$. When the twist wave vector is small compared to the plasmon wave vector at the resonance, $Q \ll q^*$, the twist-limited contribution to the tunnel current can be estimated as

$$I_{t\to b}^{\mathrm{pl,em}}(\Delta^*) \approx \frac{I_0}{2\pi}\sqrt{\frac{q^2\partial^2\omega/\partial q^2}{vQ}}\ln\left|\frac{q_c^2\partial^2\omega_q/\partial q^2}{8vQ}\right|, \qquad (20)$$

where q_c is the momentum cutoff associated with interband damping of plasmons.

Due to the complicated structure of plasmon dispersion near the critical wave vector q^*, the existence condition for plasmaronic resonance in twisted layers can be presented only in a very rough manner. In most realistic situations, the coupling constant in graphene $\alpha_c = e^2/\kappa v$ is on the order of unity; hence the numerator of the large logarithm is on the order of the Fermi wave vector k_F. Thus, the Fermi wave vector should be far above the twist vector Q, while the twist angle should satisfy $\Delta\theta \ll k_F/K_D$ where K_D is the distance between the Γ and K points in graphene. For a typical value of Fermi energy of 200 meV this yields $\Delta\theta \ll 1^\circ$; due to the weakness of logarithmic singularity, the strong inequality ($\ll$) is essential.

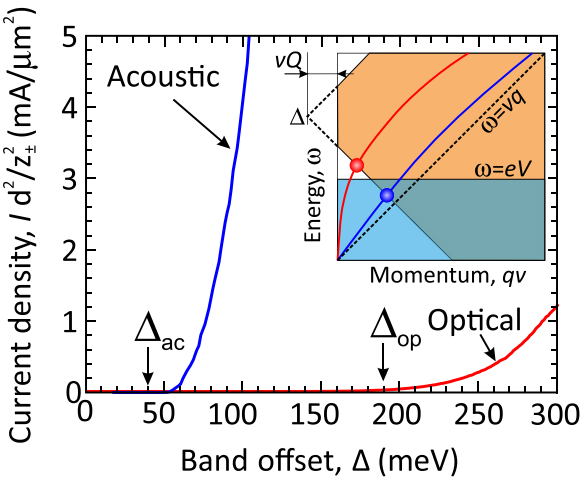

FIG. 3. Calculated dependence of plasmon-assisted tunnel current at fixed carrier densities in the layers ($\varepsilon_{F+} = 225$ meV, $\varepsilon_{F-} = 250$ meV) vs band offset Δ at $T = 0$ and perfect alignment. The offsets Δ_{ac} and Δ_{op} correspond to the switch-on of plasmon-aided tunneling with emission of acoustic and optical plasmons, respectively. An inset shows the diagram for geometrical determination of threshold voltage: the intraband tunneling with acoustic (optical) plasmon emission becomes possible when the blue (red) dot appears below the Pauli blocking line $\omega = eV$. For the conditions shown in the inset, only the emission of acoustic plasmons is possible.

B. Fine structure of the low-temperature $I(V)$ curves

Among more common manifestations of plasmon-assisted tunneling in graphene-based junctions there stands the emergence of the threshold structure in $I(V)$ curves. At low temperatures, the tunneling with plasmon absorption is frozen out, while emission-aided tunneling is possible only for $\omega_q < eV$ by virtue of Pauli blocking. The combination of the Pauli principle and energy-momentum conservation results in suppression of inelastic current for low voltages $V < V_{\mathrm{th}}$ and band offsets $\Delta < \Delta_{\mathrm{th}}$. These threshold quantities are found from

$$eV_{\mathrm{th}} = \omega_q, \quad \omega_q = \Delta_{\mathrm{th}} - v(q+Q). \qquad (21)$$

A simple geometrical interpretation of the latter system is shown in the inset of Fig. 3. The minimal frequency of plasmons in the domain of intraband tunneling $|\omega - \Delta| < v(q+Q)$ (dot at the boundary of orange filled region) should lie below the line of Pauli blocking $\omega = eV$, i.e., in blue filled region. From this analysis we also see that finite interlayer twist Q reduces the threshold of plasmon emission upon tunneling.

For acoustic plasmons in the graphene double layer with linear dispersion $\omega = sq$, the threshold condition (21) can be solved analytically to yield

$$eV_{\mathrm{th}} = \frac{\Delta - vQ}{1 + v/s}. \qquad (22)$$

If the band offset Δ is fixed, the threshold voltage (22) weakly depends on carrier density because the plasmon velocity s saturates to the Fermi velocity at small interlayer distance

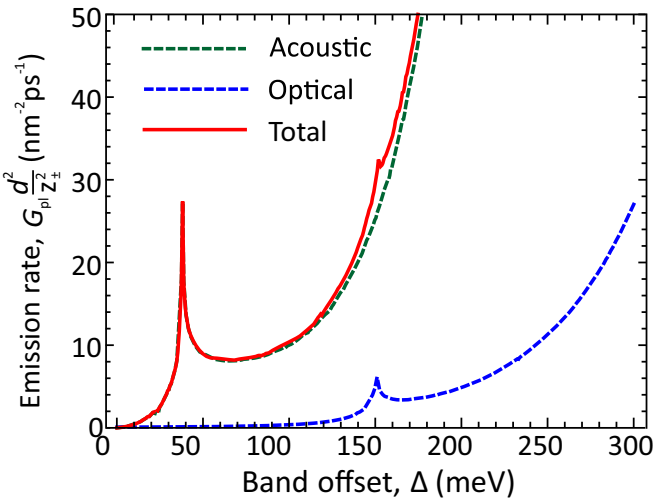

FIG. 4. Dependence of plasmon generation rate (normalized by exponential tunneling factor $z_\pm^2/d^2$) on the offset between Dirac points for the parameters corresponding to the red curve in Fig. 2(a).

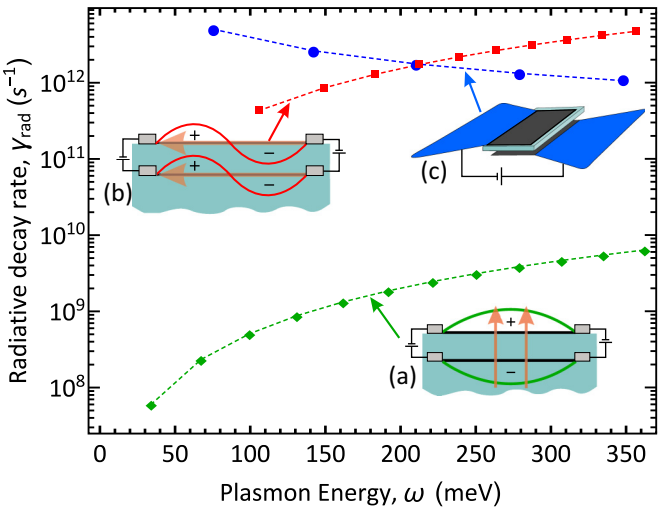

FIG. 5. Illustration of various radiative decay channels for plasmons in confined double layers and estimates of the respective radiative decay rates: (a) emission of vertical dipoles for acoustic plasmon modes, (b) emission of lateral dipoles for optical modes, and (c) emission of antenna-coupled modes. Wavy lines in (a) and (b) show the distribution of charge density in plasmon modes; arrows show the direction of the dipole moment. Dots correspond to different eigenmodes of the structures. For all curves $L = 80$ nm, $W = 1$ μm, Fermi energies $\varepsilon_{F+} = \varepsilon_{F-} = 250$ meV.

d [31,49,51]. This contrasts to the case of plasmon-assisted tunneling in bulk metal-insulator-metal junctions [25] where the threshold voltage equals the plasmon energy, the latter being proportional to the square root of density.

If the carrier densities in graphene layers are fixed while band offset is swept, the threshold condition can be presented in an alternative form:

$$\Delta_{\rm th} = \left(1 + \frac{s}{v}\right)(\varepsilon_{F-} - \varepsilon_{F+}) - sQ. \tag{23}$$

The steplike switch-on of the tunnel current upon increase in band offset Δ is shown in Fig. 3 for fixed carrier densities and zero temperature. These cusps in the $I(V)$ curves become broadened very quickly with rising temperature, as the broadening is governed both by smearing of Fermi distributions and interband damping of graphene plasmons.

C. Plasmon emission and junction electroluminescence

The emission of surface plasmons upon resonant tunneling can be detected not only by analyzing the features of inelastic current. Recent advances in near-field microscopy and nanoscale electromagnetic sensing [3,39] allow a direct measurement of plasmon generation rates. This generation rate, $G_{\rm pl}$, is obtained by a simple rearrangement of terms in the expression for plasmon-assisted tunnel current:

$$G_{\rm pl} = \frac{1}{e}\left[I_{t\to b}^{\rm pl,em} + I_{b\to t}^{\rm pl,em} - I_{t\to b}^{\rm pl,abs} - I_{b\to t}^{\rm pl,abs}\right]. \tag{24}$$

Naturally, the bias dependence of the integrated plasmon emission rate inherits all resonant features of the plasmon-assisted current. This is shown in Fig. 4, where the characteristic peaks are the plasmaronic resonances discussed above. The peak emission rate for the resonance with acoustic modes is 3×10^{27} cm^{-2} s^{-1} timed by a barrier-dependent tunneling exponent $(z_\pm/s)^2$. The barrier-independent part exceeds the typical plasmon generation rate upon interband recombination in population-inverted graphene, which is $\sim 10^{26}$ cm^{-2} s^{-1} at the same carrier density [52]. Due to the smallness of the tunneling exponent [$(z_\pm/d)^2 \approx 8 \times 10^{-7}$ for 2.5 nm of hBN barrier, 5×10^{-2} for 2.5 nm WS_2 barrier at $\Delta = 50$ meV], the real generation rate is smaller. However, compared to plasmonic oscillators based on pumped graphene [52,53], the tunneling devices are not prone to strong Auger recombination and heating.

The tunneling emission of plasmons can be also implicitly studied via analysis of tunnel junction electroluminescence that was recently observed in a sample with a pronounced interlayer twist [24]. Such electroluminescence is commonly a two-step process [25,26] including the excitation of surface plasmons upon inelastic tunneling and its subsequent radiative decay into free-space modes [54]. The fate of plasmons emitted upon tunneling is determined by competition of radiative decay (with the characteristic rate $\gamma_{\rm rad}$) and in-plane absorption due to the Drude or interband mechanism (having the rate $\gamma_{\rm sc}$). When both rates are smaller than the plasmon frequency, the number of emitted photons can be presented as

$$G_{\rm ph} = \frac{\gamma_{\rm rad}}{\gamma_{\rm rad} + \gamma_{\rm sc}} G_{\rm pl}. \tag{25}$$

The radiative decay depends on the plasmon-to-photon coupling scheme. The plasmons in confined double-layer structures possess a nonzero dipole moment which can be directed either in-plane or normally to the layers; in either case this leads to the dipole radiation. The decay rate can be estimated as $\gamma_{\rm rad} = P_{\rm rad}/w$, where $P_{\rm rad} = 4\omega^4|\mathbf{d}_\omega|^2/c^3$ is the radiation power, w is the classical mode energy, and $\mathbf{d}_\omega$ is the mode dipole moment.

The acoustic modes are prone to radiative decay via emission of vertical dipoles, shown in Fig. 5(a). The corresponding

decay rate can be shown to be (see Appendix G)

$$\gamma_{\rm rad}^{\updownarrow} = \frac{64\omega}{3}\left.\frac{(qd)^2(s/c)(\sigma''/c)^2(W/L)}{\frac{\kappa}{4\pi}[1+\coth(qd/2)] - q\frac{d\sigma''}{d\omega}}\right|_{\substack{q=\frac{\pi n}{L}\\ \omega=\omega_q^-}}, \qquad (26)$$

where W is the GIG structure width and σ'' is the imaginary part of the in-plane conductivity (for simplicity, we assume both layers to have equal carrier density). Apparently, this process is inefficient due to the smallness of plasmon phase velocity s compared to the velocity of light, the smallness of interlayer distance, and finally, the large value of mode energy in the long-wavelength limit [large denominator in Eq. (26)]. For Fermi energies $\varepsilon_{F+} = \varepsilon_{F-} = 250$ meV and structure length $L = 80$ nm we obtain the energy of the lowest acoustic mode $\omega = 33$ meV and the decay rate $\gamma_{\rm rad}^{\updownarrow} \approx 8\times 10^7$ s^{-1}; it can become somewhat larger for higher modes (green curve in Fig. 5). This is four orders of magnitude below the scattering rate in moderate-quality samples, $\gamma_{\rm sc} \approx 10^{12}$ s^{-1}.

The radiation is strongly intensified for modes possessing lateral dipole moment, shown in Fig. 5(b). For a purely optical mode in a double layer with equal in-plane conductivities, a similar estimate leads us to

$$\gamma_{\rm rad}^{\leftrightarrow} = \frac{64\omega}{3}\left.\frac{(qL)^2(\omega/cq)(\sigma''/c)^2(W/L)}{\frac{\kappa}{4\pi}[1+\tanh(qd/2)] - q\frac{d\sigma''}{d\omega}}\right|_{\substack{q=\frac{\pi n}{L}\\ \omega=\omega_q^+}}. \qquad (27)$$

This differs from the estimate for vertical dipoles of the acoustic mode by a large factor $(L/d)^2$ which comes from the difference of dipole lengths and by the smallness of the energy denominator in the $q\to 0$ limit. These two factors lead to an elevated decay rate of $\gamma_{\rm rad}^{\leftrightarrow} \approx 5\times 10^{11}$ s^{-1}; the corresponding plasmon energy is $\omega \approx 100$ meV. The radiative decay rate of this mode is comparable to the scattering rate; hence, the excitation of these modes by tunneling is readily seen in luminescence.

Probably the highest plasmon-to-photon conversion efficiency is achieved by coupling a GIG tunnel junction to a nanoantenna [27,55] [Fig. 5(c) and the blue curve]. If the antenna impedance $Z_{\rm rad}$ is matched to the impedance of the double-layer structure [$Z_{\rm rad}\sigma''(\omega_{\mathbf{q}})W/L \sim 1$], the plasmon decay rate becomes comparable to the plasma frequency. The estimates for a junction coupled to a resistive load $Z_{\rm rad}$ yield $\gamma_{\rm rad,\,max} \approx 0.04\omega$ for the lowest plasmon mode, independently of geometrical dimensions and carrier densities in the layers (for details see Appendix H). Therefore, near-unity plasmon-to-photon conversion efficiency looks possible in optimized structures.

IV. DISCUSSION AND CONCLUSIONS

We have theoretically identified several manifestations of plasmon-assisted tunneling in graphene-insulator-graphene junctions, the most striking of them being the plasmaronic resonance in tunnel current. The origin of this resonance is the enhanced interaction between plasmons and interlayer single-particle excitations due to the group velocity matching. The relation between the discussed resonance and the formation of plasmarons in a single graphene layer is elucidated as follows. Plasmarons are formed off the mass shell $\epsilon_p = pv$ at some energy separation $\delta\epsilon$ equal to the energy of plasmon quantum ω_q. Contrary to 3d systems, the plasmons in two dimensions have a soft spectrum with energy tending to zero at long wavelength. A natural question arises: Which plasmon wave vector q^* provides the strongest interaction with electrons? The answer is that such a plasmon should have group velocity equal to the carrier velocity [38]. In the case of interlayer tunneling, we are dealing essentially with on-shell electrons; however, the energy of interlayer single-particle excitations is tuned by interlayer bias Δ. At some bias Δ^*, the energies, momenta, and group velocities of interlayer excitations and plasmons coincide. This bias corresponds to the resonantly large generation of surface plasmons by interlayer tunneling.

The mentioned resonance is closely related to the square-root singularities in the joint density of states (JDOS) for the GIG structures, as discussed in Appendix D. Such singularities are inherited from the linear carrier dispersion [42] and are absent in coupled systems with parabolic bands [56]. Experimentally, the plasmonic peaks in inelastic current for coupled massive layers were observed at very specific conditions [35], e.g., at the anticrossing of intersubband plasmon dispersions [57]. The replacement of one of graphene layers with a bilayer leads to weak (logarithmic) JDOS singularities and thus finite net plasmon emission rate and inelastic current at any bias Δ.

The singularities in graphene polarizability can be modified by electron-electron corrections to carrier dispersion and/or by vertex corrections [58]. Though our original derivation of inelastic current was based on the scattering of noninteracting particles, the interaction effects can be conveniently included in the transformed equation (2) by replacing the bare polarizabilities $\Pi_{ll'}$ with interacting ones; a similar situation occurs in the theory of Coulomb drag [59]. Therefore, the expression for scattering-assisted current (2) would be valid in the vicinity of the Dirac point, where the carrier interactions are crucial. However, no plasmaronic resonances are expected in this case due to strong interband plasmon absorption. In doped samples, the interactions just enhance the band velocity under the Fermi surface [60] and the plasmaronic resonances are expected to persist.

In the present calculation, we assumed the interlayer tunneling to be weak, so that the dielectric function of the double layer was not renormalized by tunneling. Such renormalization can be done [31], and it would enhance the plasmon-assisted current. The reason for enhancement is the partial plasmon loss compensation by stimulated plasmon emission upon tunneling under interlayer population inversion. A similar stimulated tunneling process represents the principle of a quantum cascade laser [45] and was also proposed for amplification of plasmons [31,61]. At some critical strength of tunneling, corresponding to the complete undamping of plasmon modes, the current (2) would diverge. The divergence would signal the onset of surface plasmon lasing; at this point one has to solve the coupled kinetic equations for electrons and plasmons for evaluation of tunnel current.

The present theory demonstrates the prospect of graphene heterostructures for resonant and voltage tunable light emission in the far infrared. Compared to the light sources based on carrier injection and recombination (in transition metal dichalcogenides [7] and graphene quantum dots [62]), the proposed structures offer voltage tunability of the emission spectrum, and can be integrated in photonic [9] and plasmonic

[63] waveguides. The proposed process of plasmon and photon generation is different from interband recombination of electrons injected upon resonant tunneling considered in [64]. The difference between these processes is the same as the difference of quantum-cascade lasing with vertical and diagonal radiative transitions. Importantly, the emission spectrum for interband recombination of injected carriers is smooth, following the interband JDOS spectrum.

In conclusion, we have developed a theoretical formalism for the calculation of tunneling current accompanied by carrier-carrier scattering in graphene-insulator-graphene heterostructures. Our calculation shows that the main contribution to inelastic scattering-assisted current comes from emission of surface plasmons. The plasmon-assisted current can be resonantly enhanced if the energy, momentum, and group velocity of interlayer excitations and plasmons coincide. This effect, which we call plasmaronic resonance, can also manifest itself in enhanced plasmon emission and electroluminescence of graphene-based junctions.

ACKNOWLEDGMENTS

This work was supported by Grants No. 16-37-60110/16 and No. 16-29-03402/16 of the Russian Foundation for Basic Research and by the grant of the president of the Russian Federation No. SP-589.2016.5. The authors are grateful to G. Alymov for valuable discussions.

APPENDIX A: PLASMON POLES IN SCATTERING-ASSISTED TUNNELING

The current accompanied by emission of plasmons can be derived by extracting the contribution to the integral (2) due to the poles of the screening function $\varepsilon^{-1}(\mathbf{q},\omega)$. Assuming the dissipation of electromagnetic energy to be small, one can determine the plasmon frequency $\omega_{\mathbf{q}}^{p}$ from

$$\varepsilon'\left(\mathbf{q},\omega_{\mathbf{q}}^{p}\right)=0. \tag{A1}$$

If the frequency ω and momentum $\mathbf{q}$ satisfy the dispersion relation (A1), the transition amplitudes $V_{++,-+}$ and $V_{+-,--}$ are related as follows:

$$V_{++,-+}=V_{+-,--}e^{-qd}/S_{-}. \tag{A2}$$

With the help of Eqs. (A1) and (A2) we can write down the tunneling current [Eq. (2) of the main text] as follows (keeping in mind $\omega\approx\omega_{\mathbf{q}}$):

$$I_{t\to b}=\frac{2e}{\pi}\int_{-\infty}^{+\infty}d\omega\sum_{\mathbf{q},l}\Pi''_{+-}(\mathbf{q},\omega)N_{\omega-eV}(N_{\omega}+1)\times\frac{|V_{+-,--}|^{2}}{|S_{-}|}[-\varepsilon''(\mathbf{q},\omega)]. \tag{A3}$$

When the damping rate of plasmons is small (i.e., $|\varepsilon''/\varepsilon'|\ll 1$) one can single out the plasmon-assisted contribution

$$\frac{\varepsilon''(\mathbf{q},\omega)}{|\varepsilon(\mathbf{q},\omega)|^{2}}\approx 2\pi\sum_{p=\pm1}\frac{\delta\left(\omega-\omega_{q}^{p}\right)+\delta\left(\omega+\omega_{q}^{p}\right)}{|\partial\varepsilon'/\partial\omega|_{\omega_{q}^{p}}} \tag{A4}$$

in Eq. (A3). Then, Coulomb interaction between two electrons is reduced to $|V_{+-.--}|=V_{0}(\mathbf{q})|S_{\mathbf{q}}(z)/\varepsilon(\mathbf{q},\omega_{\mathbf{q}}^{p})|$ and we obtain the resulting expression for the tunneling current:

$$I_{t\to b}=4e\sum_{\mathbf{q},p=\pm1}\frac{V_{0}(\mathbf{q})|S_{\mathbf{q}}(z)|^{2}}{|S_{-}\partial\varepsilon'/\partial\omega|_{\omega_{\mathbf{q}}^{p}}}\times\left[\Pi''_{+-}\left(\mathbf{q},\omega_{\mathbf{q}}^{p}\right)N_{\omega_{\mathbf{q}}^{p}-eV}\left(N_{\omega_{\mathbf{q}}^{p}}+1\right)+\Pi''_{-+}\left(\mathbf{q},\omega_{\mathbf{q}}^{p}\right)N_{\omega_{\mathbf{q}}^{p}}\left(N_{\omega_{\mathbf{q}}^{p}+eV}+1\right)\right]. \tag{A5}$$

The first term in square brackets of Eq. (A5) is nothing but the tunneling accompanied by emission of plasmons (11) and the second one by absorption (12).

APPENDIX B: ENERGY OF FIELD IN PLASMONIC MODES

In this section we find the energy of electromagnetic field in the plasmon modes coupled to the graphene double layer. The distribution of the electric potential is harmonic with respect to in-plane coordinates $\mathbf{r}_{\parallel}$:

$$\varphi(\mathbf{r}_{\parallel},z)=\frac{1}{2}\sum_{\mathbf{q}}\varphi_{\mathbf{q}}(z)e^{i\mathbf{q}\mathbf{r}_{\parallel}-i\omega t}+\text{c.c.} \tag{B1}$$

The Fourier components $\varphi_{\mathbf{q}}(z)$ satisfy the Poisson equation

$$\frac{\partial^{2}\varphi_{\mathbf{q}}}{\partial z^{2}}-q^{2}\varphi_{\mathbf{q}}=-4\pi e[\rho_{+}\delta(z-d/2)+\rho_{-}\delta(z+d/2)], \tag{B2}$$

where $\rho_{\pm}$ are the induced charge densities in the top and bottom layers. We supplement Eq. (B2) with the relation between induced charge density and potential:

$$\rho_{\pm}=e^{2}\Pi_{\pm}(\mathbf{q},\omega)\varphi_{\mathbf{q}}(z=\pm d/2). \tag{B3}$$

The solution for guided modes is parametrized as $\varphi_{\mathbf{q}}(z)=\varphi_{0}S_{\mathbf{q}}(z)$, while the spatial dependence is given by

$$S_{\mathbf{q}}(z)=\begin{cases}S_{+}e^{-q(z-d/2)}, & z>d/2,\\ \frac{S_{+}\sinh\left[q\left(z+\frac{d}{2}\right)\right]-S_{-}\sinh\left[q\left(z-\frac{d}{2}\right)\right]}{\sinh qd}, & |z|\leqslant d/2,\\ S_{-}e^{q(z+d/2)}, & z<-d/2,\end{cases} \tag{B4}$$

with $S_{+}=e^{-qd}$, $S_{-}=1-V_{0}(\mathbf{q})\Pi_{++}(\mathbf{q},\omega)(1-e^{-2qd})$. Plugging Eq. (B4) into the Brillouin formula (14), we find the energy stored in the plasmon mode:

$$w=\frac{A\varphi_{0}^{2}q\omega_{q}S_{-}}{8\pi}\left.\frac{\partial\varepsilon}{\partial\omega}\right|_{\omega=\omega_{q}^{p}}. \tag{B5}$$

Equating this energy to ω_{q}, we find the zero-point amplitude of the electric potential:

$$\left(\frac{e\varphi_{0}}{2}\right)^{2}=\frac{V_{0}(\mathbf{q})}{AS_{-}\,\partial\varepsilon'/\partial\omega|_{\omega=\omega_{q}^{p}}}. \tag{B6}$$

APPENDIX C: SINGULAR STRUCTURE OF POLARIZABILITY AND PLASMARONIC RESONANCE

In this section, we discuss how the singular structure of interlayer polarizability is related to the resonant features of inelastic current. Extracting the imaginary part of polarizability [Eq. (3)] with the aid of the Sokhotski theorem, one can

obtain the following expression:

$$\Pi''_{+-}(\mathbf{q},\omega) = \frac{\pi g q^2}{2(2\pi)^2}\sum_{s=\pm1}\left[\frac{\theta(vq-|\omega-\Delta|)}{\sqrt{(vq)^2-(\omega-\Delta)^2}}J_{1,s}+\frac{\theta(|\omega-\Delta|-vq)}{\sqrt{(\omega-\Delta)^2-(vq)^2}}J_{2,s}\right]$$
$$\equiv \frac{q^2}{2\pi}\left[\frac{\widetilde{\Pi}_{\pm,1}(\mathbf{q},\omega)}{\sqrt{(vq)^2-(\omega-\Delta)^2}}+\frac{\widetilde{\Pi}_{\pm,2}(\mathbf{q},\omega)}{\sqrt{(\omega-\Delta)^2-(vq)^2}}\right]. \quad \text{(C1)}$$

The first term in the square brackets is due to intraband and the second one to the interband tunneling. The last line is the definition of the nonsingular part of the polarizability $\widetilde{\Pi}_{\pm}$. The integrals $I_{1,s}$ and $I_{2,s}$ involve electron distribution functions in the opposite layers:

$$J_{1,s} = \int_1^{+\infty} dx\sqrt{x^2-1}\left[f^+\left(\frac{svqx+\omega-\Delta}{2}\right)-f^-\left(\frac{svqx-\omega+\Delta}{2}\right)\right],$$
$$J_{2,s} = \int_0^{1} dx\sqrt{1-x^2}\left[f^+\left(\frac{\omega-\Delta-svqx}{2}\right)-f^-\left(\frac{-\omega-\Delta-svqx}{2}\right)\right].$$

Now we show how the presence of square-root singularities in Eq. (C1) results in resonance in the plasmon-assisted tunneling current (17). We assume that band offset is in the vicinity of the value $\Delta^* = vq^* + \omega^p_{q^*}$ and $\Delta < \Delta^*$. Then, we come to the following formula for the current:

$$I^{pl}_{t\to b} = 2\pi e\int_0^{+\infty}\frac{qdq}{2\pi}\left|\frac{e\varphi^p_{\mathbf{q}\pm}}{2}\right|^2\left[N_{\omega^p_{\mathbf{q}}}+1\right]N_{\omega^p_{\mathbf{q}}-eV}\frac{q^2}{2\pi}\frac{\widetilde{\Pi}_{\pm,1}\left(\mathbf{q},\omega^p_{\mathbf{q}}\right)}{\sqrt{(vq)^2-\left(\omega^p_{\mathbf{q}}-\Delta\right)^2}}. \quad \text{(C2)}$$

We expand the plasmon dispersion in the vicinity of $q = q^*$ that corresponds to the tangent of the dispersion curve and single-particle excitation threshold, $\omega_q \approx \omega_{q^*} + v(q-q^*) + \partial^2\omega_q/\partial q^2(q-q^*)^2/2$. Evaluating the nonsingular terms at $q = q^*$, $\omega = \omega_{q^*}$, we find the leading behavior of the plasmon-assisted current:

$$I^{pl}_{t\to b} \approx \left|\frac{e\varphi_{\mathbf{q}\pm}}{2}\right|^2\frac{eq^3[N_\omega+1]N_{\omega-eV}\widetilde{\Pi}_{\pm,1}(\mathbf{q},\omega)}{2\pi\sqrt{vq\partial^2\omega_{\mathbf{q}}/\partial q^2}}\int_{-x_0}^{x_0}\frac{dx}{\sqrt{1+x^2}} \approx \left|\frac{e\varphi_{\mathbf{q}\pm}}{2}\right|^2\frac{eq^3[N_\omega+1]N_{\omega-eV}\widetilde{\Pi}_{\pm,1}(\mathbf{q},\omega)}{2\pi\sqrt{vq\partial^2\omega_{\mathbf{q}}/\partial q^2}}\ln x_0^2, \quad \text{(C3)}$$

where

$$x_0^2 = \frac{q^2\partial^2\omega_{\mathbf{q}}/\partial q^2}{2(\Delta-\Delta^*)}. \quad \text{(C4)}$$

APPENDIX D: JOINT DENSITY OF STATES AND ANOMALIES IN TUNNELING SPECTRA

The singular structure of the tunneling spectra can be analyzed by calculating the joint density of states (JDOS) for optical transitions between two coupled layers. Indeed, the singularity in the interlayer polarizability (C1) is nothing but the singularity in the JDOS occurring at $\omega = \Delta + qv$. In general, the JDOS can be introduced for optical transitions between two-dimensional systems with carrier dispersions $\epsilon^+_{\mathbf{p}}$ and $\epsilon^-_{\mathbf{p}}$:

$$D(\mathbf{q},\omega) = \sum_{\mathbf{p}}\delta(\epsilon^+_{\mathbf{p}}+\Delta-\omega-\epsilon^-_{\mathbf{p}-\mathbf{q}}), \quad \text{(D1)}$$

where the summation is carried over all occupied states in the top (+) layer and unoccupied states in the bottom (−) one. It is easy to see that for coupled single layers of graphene $\epsilon^+_{\mathbf{p}} = \epsilon^-_{\mathbf{p}} = pv$, and the JDOS acquires a square-root singularity $D(\mathbf{q},\omega) \propto [(qv)^2-(\Delta-\omega)^2]$.

It is possible to consider the spectrum of inelastic tunneling between the single graphene layer and two-dimensional system with parabolic bands, $\varepsilon^-_{\mathbf{p}} = p^2/2m^*$. The latter spectrum describes the electron states in the graphene bilayer in the absence of transverse electric field. Integrating over the angle between $\mathbf{p}$ and $\mathbf{q}$ we arrive at the following expression for the JDOS:

$$D(\mathbf{q},\omega) \propto \int\frac{pdp}{\sqrt{F(p)}}, \quad \text{(D2)}$$
$$F(p) = [2pq]^2-[2m^*(pv+\Delta-\omega)-(p^2+q^2)]^2. \quad \text{(D3)}$$

An anomaly in the JDOS can occur provided $\Delta-\omega < m^*v^2/2$. In this regime there exist two roots of the equation $F(p) = 0$,

$$p_\pm = m^*v-q\pm\sqrt{2m(\Delta-\omega-qv)-m^2v^2}, \quad \text{(D4)}$$

merging at $qv \to mv^2/2+\Delta-\omega$. The anomaly in the JDOS at such momentum transfer is logarithmic. The result of the numerical calculation of the JDOS with Eq. (D2) is shown in Fig. 6 for three different values of band offset $\Delta-\omega$ and effective mass appropriate to the graphene bilayer, $mv^2 \approx 200$ meV.

Though the logarithmic singularities in JDOS between single layer and bilayer graphene can be observed in the energy spectra of emitted plasmons, the net plasmon emission rate (and net elastic current) remain finite and smooth functions of the band offset Δ. The reason is that singularity in the JDOS in such a system is only logarithmic, and its integration over wave vectors of emitted plasmons q would lead to finite net emission.

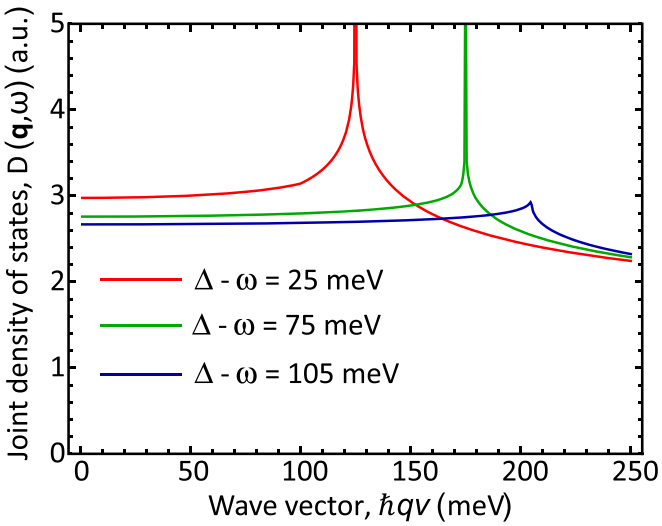

FIG. 6. Joint density of states for tunneling transitions between single-layer graphene and 2d electron system with parabolic dispersion at various band offsets Δ. Singularities in the red and green curves occur at $qv = mv^2/2 + \Delta - \omega$.

Using a similar method, it is possible to show that the JDOS for tunneling between two massive two-dimensional electron systems has no singularities except for the case of photon-assisted tunneling ($q = 0$).

APPENDIX E: TUNNELING MATRIX ELEMENTS

Electron states in coupled graphene layers with small interlayer twist can be described using the Hamiltonian

$$\hat{H}_0 = \begin{pmatrix} \hat{H}_{G+} & \hat{\mathcal{T}} \\ \hat{\mathcal{T}}^* & \hat{H}_{G-} \end{pmatrix}, \tag{E1}$$

where the blocks $\hat{H}_{G\pm}$ describe isolated graphene layers, and $\hat{\mathcal{T}}$ describes tunnel hopping:

$$\hat{\mathcal{T}} = \frac{\Omega}{3} \sum_{j=0,1,2} e^{-i\mathbf{Q}_j\mathbf{r}} \begin{pmatrix} 1 & e^{-i\frac{2\pi j}{3}} \\ e^{i\frac{2\pi j}{3}} & 1 \end{pmatrix}. \tag{E2}$$

Here Ω is the tunneling overlap integral and $\mathbf{Q}_j$ are the wave vectors connecting the respective edges of the hexagonal Brillouin zones in the layers. In the absence of the interlayer twist ($\mathbf{Q}_j = 0$), the tunneling matrix is diagonal, $\hat{\mathcal{T}} = \Omega\hat{I}$, where $\hat{I}$ is the identity matrix. In this case, the band and layer degrees of freedom are decoupled.

We proceed now to the evaluation of matrix element Ω. The physical meaning of Ω is half the energy splitting between electron states in coupled graphene layers, as can be seen from diagonalization of Hamiltonian (E1). On the other hand, this splitting can be estimated from a continuum model, where each graphene layer is represented by a delta well [31]. The delta-well potential is

$$U(z) = 2\sqrt{\frac{\hbar^2 U_0}{2m^*}}[\delta(z - d/2) + \delta(z + d/2)], \tag{E3}$$

where U_0 is the work function from graphene to the barrier material, and m^* is the effective mass in the barrier. For boron nitride, $U_0 \approx 1.5$ eV and $m^* \approx 0.5m_0$. The eigenfunctions in this potential are symmetric and antisymmetric ones. The energy difference between these states is

$$E_+ - E_- = 2\Omega = 4U_0 e^{-\varkappa d}, \tag{E4}$$

where $\varkappa = \sqrt{2m^*U_0}/\hbar$ is the decay constant of the electron wave function.

The plasmon-assisted current is proportional to the matrix element of electric potential energy $e\varphi_\pm$. Its evaluation generally requires the knowledge of the electron wave function inside the barrier layer. This evaluation can be, however, simplified in the dipole approximation. We write the potential distribution in the plasmon mode as $\varphi_q(z) = \bar{\varphi} + (\varphi_+ - \varphi_-)z/d$, thus the potential matrix element becomes

$$e\varphi_\pm \approx (\varphi_+ - \varphi_-)\frac{z_\pm}{d}. \tag{E5}$$

It appears that the coordinate matrix element $z_\pm$ and the tunnel splitting Ω are bound by a simple relation. We consider two methods for calculation of current between states $|+\rangle$ and $|-\rangle$. On one hand, this can be expressed through velocity operator in the transverse direction:

$$j_\pm = \frac{(v_z)_\pm}{d} = \frac{z_\pm}{\hbar d}(\epsilon_+ - \epsilon_-) = \frac{z_\pm}{\hbar d}\sqrt{\Delta^2 + 4\Omega^2}. \tag{E6}$$

On the other hand, it can be found by evaluating the derivative of the particle number in the state $|+\rangle$,

$$j_\pm = \frac{dN_+}{dt} = -\frac{i}{\hbar}[\hat{N}_+, \hat{H}_0] = \frac{\Omega}{\hbar}. \tag{E7}$$

Comparing these two expressions, we find

$$z_\pm = d\frac{\Omega}{\sqrt{\Delta^2 + 4\Omega^2}}. \tag{E8}$$

APPENDIX F: COMPARISON OF ELASTIC AND INELASTIC CURRENTS

The elastic current can be evaluated by considering the tunneling matrix elements in Hamiltonian (E1) as small perturbations. This leads to the following formula [20]:

$$I^{el} = \frac{ge}{\hbar}\sum_{\mathbf{k}ss'}\int_{-\infty}^{+\infty}\frac{dE}{2\pi}\left|\Omega^{+s,-s'}_{\mathbf{k},\mathbf{k}+\mathbf{Q}}\right|^2 \times A_{+s}(\mathbf{k},E)A_{-s'}(\mathbf{k}+\mathbf{Q},E)[f(E) - f(E - eV)]; \tag{F1}$$

here $A_{ls}(\mathbf{k},E) = -2\mathrm{Im}G^R_{ls}(\mathbf{k},E)$ is the spectral function in the lth layer and sth band, and $G^R_{ls}(\mathbf{k},E)$ is the retarded Green's function in graphene in the band representation:

$$G^R_{ls}(\mathbf{k},E) = \left[E - \epsilon^{ls}_{\mathbf{k}} - \Sigma^{ls}(\mathbf{k},E) + i\delta\right]^{-1}; \tag{F2}$$

$\Sigma^{ls}(\mathbf{k},E)$ is the electron self-energy. In the simplest approximation, the self-energy can be treated as a constant $\Sigma^{ls}(\mathbf{k},E) \approx \gamma$. In this case, one can approximate

$$I^{el} \approx \frac{2\pi ge}{\hbar}\sum_{\mathbf{k}ss'}\left|\Omega^{+s,-s'}_{\mathbf{k},\mathbf{k}+\mathbf{Q}}\right|^2\delta_\gamma\left(\epsilon^{+s}_{\mathbf{k}} - \epsilon^{-s'}_{\mathbf{k}+\mathbf{Q}}\right)\left[f^{+s}_{\mathbf{k}} - f^{-s'}_{\mathbf{k}+\mathbf{Q}}\right], \tag{F3}$$

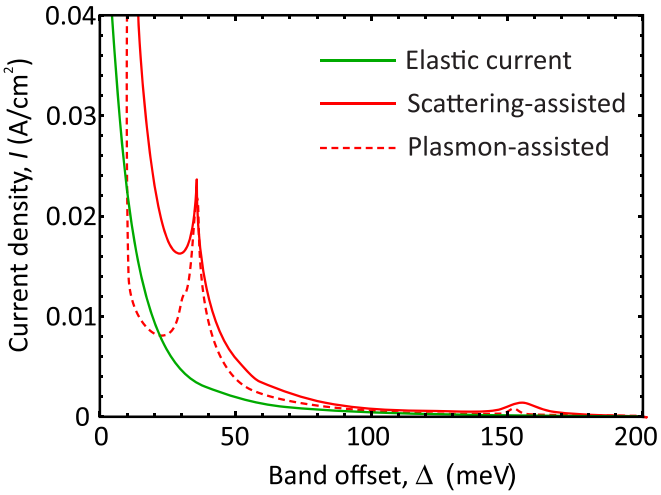

FIG. 7. Dependence of absolute values of tunnel currents on band offset at $\varepsilon_{F,+} = 0.6$ eV, $\varepsilon_{F,-} = -0.2$ eV, $d = 38$ Å. We use broadening $\gamma = 10$ meV for the elastic current.

where $\delta_\gamma(x) = (\gamma/\pi)/(\gamma^2 + x^2)$ is the "broadened" delta function. If, in addition, the layers are aligned ($\mathbf{Q} \to 0$), the integration is performed trivially yielding the electron (n) and hole (p) densities in the layers:

$$I^{el} \approx \frac{2\pi e}{\hbar}|\Omega|^2\delta_\gamma(\Delta)[(n_+ - p_+) - (n_- - p_-)]. \quad \text{(F4)}$$

The comparison of elastic and inelastic currents in aligned layers is shown in Fig. 7, where the red solid curve is obtained using Eq. (2) for scattering-aided tunneling and the red dashed curve in the plasmon pole approximation. Both elastic and inelastic components go to zero with increasing band offset, the latter due to reduction in the dipole matrix element at large level spacing, $z_\pm \approx d\Omega/\Delta$. Despite this fact, both optical and acoustic peaks are clearly visible, while the inelastic component exceeds the elastic at the resonance by a factor of 3. Figure 7 is obtained for interlayer spacing of 3.8 nm, but we do not expect significant changes at other barrier thicknesses as both components of the current have the same tunneling smallness factor $(\Omega/\Delta)^2$.

APPENDIX G: RADIATIVE DECAY OF PLASMON MODES

The present section is aimed at the estimate of plasmon-to-photon conversion efficiency. The plasmon emitted upon tunneling can decay either radiatively (with the rate $\gamma_{\rm rad}$) or it can be reabsorbed due to the Drude or intraband absorption in a single layer (the corresponding rate is $\gamma_{\rm sc}$). Considering these competing channels of plasmon decay, we can estimate the plasmon-to-photon conversion probability as $\gamma_{\rm rad}/(\gamma_{\rm rad} + \gamma_{\rm sc})$.

The radiative decay rate of plasmon $\gamma_{\rm rad}$ can be estimated as

$$\gamma_{\rm rad} = \frac{P_{\rm rad}}{w}, \quad \text{(G1)}$$

where $P_{\rm rad} = (4\omega^4/3c^3)|\mathbf{d}_\omega|^2$ is the power of dipole radiation, and w is the mode energy calculated previously [Eq. (14)]. We consider a double graphene layer sample of length L and width W, so that L corresponds to the fundamental plasmon mode $qL = \pi$. In a general situation when the conductivities of top and bottom layers are not equal, one can evaluate the power emitted by vertical dipoles as

$$P^{\updownarrow}_{\rm rad} = \frac{4\omega^2}{3c^3}\varphi_0^2(qdW)^2\left[\sigma''_+ S_{\mathbf{q}}\left(\frac{d}{2}\right) - \sigma''_- S_{\mathbf{q}}\left(-\frac{d}{2}\right)\right]^2, \quad \text{(G2)}$$

where the dimensionless potential profile $S_{\mathbf{q}}(z)$ was introduced in Eq. (B4). Similarly, for lateral dipoles we find

$$P^{\leftrightarrow}_{\rm rad} = \frac{4\omega^2}{3c^3}\varphi_0^2(qLW)^2\left[\sigma''_+ S_{\mathbf{q}}\left(\frac{d}{2}\right) + \sigma''_- S_{\mathbf{q}}\left(-\frac{d}{2}\right)\right]^2. \quad \text{(G3)}$$

Combining the expressions for the radiated power (G2) and (G3) with the expression for mode energy (14), we find the radiative decay rates of the respective modes (26) and (27).

APPENDIX H: ANTENNA COUPLING OF PLASMONS

The plasmon-to-photon conversion efficiency can be markedly increased if the double-layer device is loaded with an antenna. To model the plasmon decay in this situation, we consider the two graphene layers connected via the radiative resistance $Z_{\rm rad}$. We shall solve the dispersion equation for plasmons in this structure and find their decay rate due to radiation. An electric potential is sought for as a superposition of forward and backward optical and acoustic waves:

$$\varphi_\pm = ae^{iq_+x} + be^{-iq_+x} \pm ce^{iq_-x} \pm de^{-iq_-x}, \quad \text{(H1)}$$

where q_- and q_+ are the wave vectors of acoustic and optical modes. The boundary conditions for the schematic in Fig. 5(c) are

$$\left.\frac{\partial\varphi_+}{\partial x}\right|_{L/2} = \left.\frac{\partial\varphi_-}{\partial x}\right|_{-L/2} = 0, \quad \text{(H2)}$$

$$\varphi_+|_{-L/2} = -\varphi_-|_{L/2} = \tfrac{1}{2}IZ_{\rm rad}, \quad \text{(H3)}$$

where I is the current induced in external circuit. Solving Eq. (H1) with boundary conditions (H2), we obtain the following dispersion relation:

$$1 + \cos q_+L\cos q_-L - \frac{1}{2}\left[\frac{q_+}{q_-} + \frac{q_-}{q_+}\right]\sin q_-L\sin q_+L = \sigma Z_{\rm rad}\frac{W}{L}\left[\frac{q_-}{q_+}\sin q_+L\sin q_-L - \sin^2\frac{q_-L}{2}\sin^2\frac{q_+L}{2}\right]. \quad \text{(H4)}$$

It is possible to estimate the solutions analytically in the limit $q_+/q_- \ll 1$; i.e., when the wavelength of the optical plasmon much exceeds that of the acoustic one. This is generally fulfilled as the acoustic plasmons have linear dispersion while the optical have a square-root one [see also Fig. 2(b)]. In this limit, the general dispersion equation (H4) is decoupled into two, neither depending on q_+:

$$\cos\frac{q_-L}{2} = 0, \quad \text{(H5)}$$

$$\frac{q_-L}{2}\tan\frac{q_-L}{2} = \frac{1}{1 + 2\sigma Z_{\rm rad}W/L}. \quad \text{(H6)}$$

Only the solutions of the second equations are affected by the radiative decay. It is convenient to rewrite it introducing

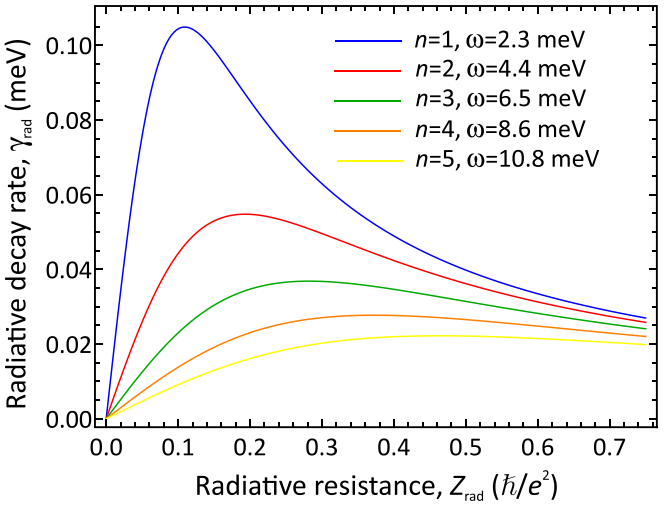

FIG. 8. Radiative decay rate of plasmon modes supported by the double graphene layer vs antenna radiative resistance $Z_{\rm rad}$ (measured in units of $\hbar/e^2$). Channel length $L = 2\ \mu$m, channel width $W = L$, interlayer distance $d = 2.5$ nm, Fermi energy $\varepsilon_{F+} = \varepsilon_{F-} = 75$ meV, $T = 300$ K.

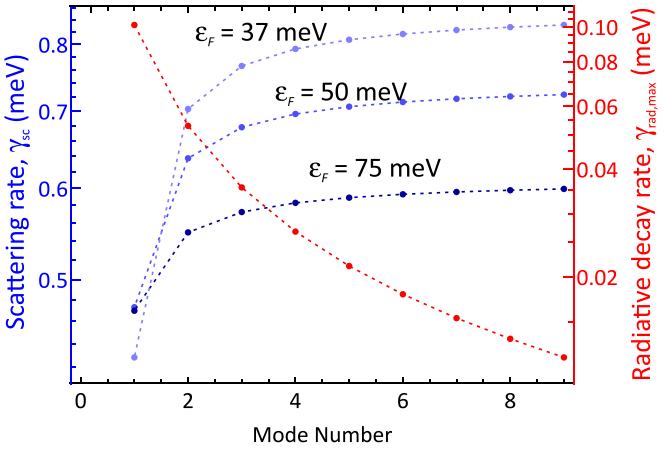

FIG. 9. Comparison of radiative ($\gamma_{\rm rad}$) and scattering (γ_{sc}) decay channels for plasmons supported by the double layer (structure parameters as in Fig. 8) at different Fermi energies. $\gamma_{\rm rad}$ is calculated for antenna resistance $Z_{\rm rad}$ providing the maximum radiative decay. The scattering rate γ_{sc} is limited by graphene acoustic phonons and residual charge impurities with density $N_i = 10^{11}$ cm^{-2}. Dashed lines are a guide for the eye.

the dimensionless frequency $u = q_- L/2 = \omega L/2s$, and the dimensionless radiative resistance

$$\tilde{Z} = Z_{\rm rad}\frac{W}{L}\frac{e^2}{\hbar}\frac{\varepsilon_F}{\hbar\omega_{pl}}, \tag{H7}$$

where $\omega_{pl} = \pi s/L$. The dispersion equation becomes

$$u\tan u = \frac{1}{1 - i\tilde{Z}/u}. \tag{H8}$$

A general feature of its solutions is that the imaginary part of the frequency has an extremum as a function of $\tilde{Z}$. This is illustrated in Fig. 8 which shows the decay rate of five lowest plasmon modes vs radiative resistance $Z_{\rm rad}$ calculated with numerical solution of Eq. (H6).

There exists an optimal value of antenna resistance providing the maximum radiative decay rate. Decoupling the solutions into real and imaginary parts, $u = u' + iu''$, we find that the maximum of the decay rate is achieved if

$$\tilde{Z} = u', \tag{H9}$$

$$u'' \approx \frac{\cos^2 u'}{2u'}. \tag{H10}$$

For the two lowest modes we have obtained $u' = 3.4$, $u'' = -0.14$ and $u' = 6.4$, $u'' = -0.08$, respectively. We note that the condition of maximum radiative decay $\tilde{Z} = u' \sim 1$ represents the matching of antenna impedance and impedance of the graphene layer at the resonant plasmon frequency. For the lowest mode, the maximum decay rate is $\gamma_{\rm rad} \approx 0.04\omega$. This greatly exceeds the decay rate due to dipole radiation into free space.

Finally, we estimate the rate of plasmon absorption due to the Drude loss. It is worthwhile noting that $\gamma_{\rm sc}$ is not just the inverse of the electron momentum relaxation rate due to the non-negligible spatial dispersion of conductivity [65]. To account for the spatial dispersion and electron scattering simultaneously, one can solve the kinetic equation for electrons with the particle-conserving collision integral [31]. This leads

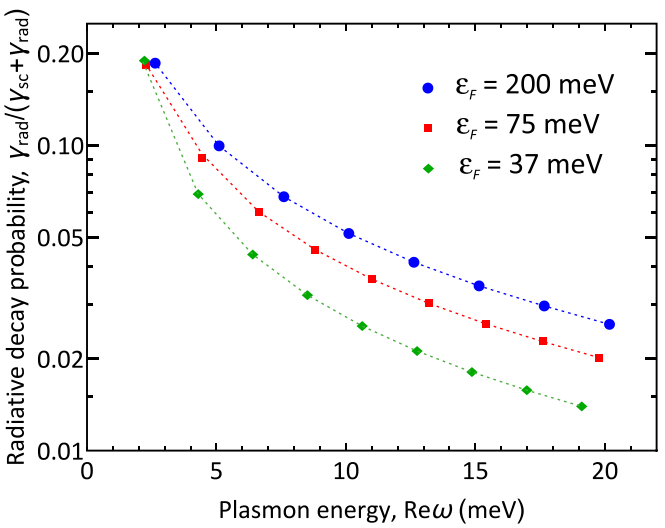

FIG. 10. Probability of plasmon decay into free space modes $\gamma_{\rm rad}/(\gamma_{\rm rad} + \gamma_{sc})$ vs frequency for different Fermi energies in the layers. Dashed lines are a guide for the eye.

to the following expression for in-plane conductivity:

$$\sigma = i\frac{g}{2\pi}\frac{e^2}{\hbar}\frac{\varepsilon_F}{qv}x\left\{\frac{x/\sqrt{x^2-1}-1}{1-(i\nu/\omega)[x/\sqrt{x^2-1}-1]}\right\}, \quad \text{(H11)}$$

where $x = (\omega + i\nu)/qv$, and ν is the electron scattering rate. Considering electron-phonon and electron-impurity collisions as the dominant scattering sources, we evaluate ν as [66,67]

$$\nu = \frac{\varepsilon_F}{T}\frac{D^2T^2}{4\rho c_s^2 v^2} + \frac{\pi}{16}\frac{v^2 N_i}{\varepsilon_F}J(\alpha_c). \quad \text{(H12)}$$

Here $D \approx 30$ eV is the deformation potential in graphene, $\rho = 7.6 \times 10^{-7}$ kg/m^2 is its mass density, $c_s = 2 \times 10^4$ m/s is the sound velocity, N_i is the impurity density, and $J(\alpha_c)$ is the dimensionless integral

$$J(\alpha_c) = \int_0^{2\pi} \frac{d\theta(1-\cos^2\theta)}{[1+(2\alpha_c)^{-1}\sin(\theta/2)]^2}. \quad \text{(H13)}$$

The results of the scattering rate and conversion efficiency calculations are shown in Figs. 9 and 10, respectively. An increase in scattering rate with reducing the Fermi energy in Fig. 9 is due to the impurity scattering contribution to plasmon damping which scales as ε_F^{-1}. At lower impurity density, the scattering will be dominated by phonons and an increase in Fermi energy would increase the scattering rates.

[1] A. K. Geim and I. V. Grigorieva, Nature (London) **499**, 419 (2013).

[2] J. R. Wallbank, D. Ghazaryan, A. Misra, Y. Cao, J. S. Tu, B. A. Piot, M. Potemski, S. Pezzini, S. Wiedmann, U. Zeitler, T. L. M. Lane, S. V. Morozov, M. T. Greenaway, L. Eaves, A. K. Geim, V. I. Fal'ko, K. S. Novoselov, and A. Mishchenko, Science **353**, 575 (2016).

[3] A. Woessner, M. B. Lundeberg, Y. Gao, A. Principi, P. Alonso-González, M. Carrega, K. Watanabe, T. Taniguchi, G. Vignale, M. Polini, J. Hone, R. Hillenbrand, and F. H. L. Koppens, Nat. Mater. **14**, 421 (2015).

[4] Q. Ma, T. I. Andersen, N. L. Nair, N. M. Gabor, M. Massicotte, C. H. Lui, A. F. Young, W. Fang, K. Watanabe, T. Taniguchi, J. Kong, N. Gedik, F. H. L. Koppens, and P. Jarillo-Herrero, Nat. Phys. **12**, 455 (2016).

[5] F. Koppens, T. Mueller, P. Avouris, A. Ferrari, M. Vitiello, and M. Polini, Nat. Nanotechnol. **9**, 780 (2014).

[6] M. Massicotte, P. Schmidt, F. Vialla, K. G. Schädler, A. Reserbat-Plantey, K. Watanabe, T. Taniguchi, K.-J. Tielrooij, and F. H. Koppens, Nat. Nanotechnol. **11**, 42 (2016).

[7] F. Withers, O. Del Pozo-Zamudio, S. Schwarz, S. Dufferwiel, P. M. Walker, T. Godde, A. P. Rooney, A. Gholinia, C. R. Woods, P. Blake, S. J. Haigh, K. Watanabe, T. Taniguchi, I. L. Aleiner, A. K. Geim, V. I. Falko, A. I. Tartakovskii, and K. S. Novoselov, Nano Lett. **15**, 8223 (2015).

[8] V. Ryzhii, T. Otsuji, M. Ryzhii, V. Y. Aleshkin, A. A. Dubinov, D. Svintsov, V. Mitin, and M. S. Shur, 2D Mater. **2**, 025002 (2015).

[9] M. Liu, X. Yin, and X. Zhang, Nano Lett. **12**, 1482 (2012).

[10] D. Svintsov, V. Vyurkov, V. Ryzhii, and T. Otsuji, J. Appl. Phys. **113**, 053701 (2013).

[11] B. N. Narozhny and A. Levchenko, Rev. Mod. Phys. **88**, 025003 (2016).

[12] R. V. Gorbachev, A. K. Geim, M. I. Katsnelson, K. S. Novoselov, T. Tudorovskiy, I. V. Grigorieva, A. H. MacDonald, S. V. Morozov, K. Watanabe, T. Taniguchi, and L. A. Ponomarenko, Nat. Phys. **8**, 896 (2012).

[13] S. Kim, I. Jo, J. Nah, Z. Yao, S. K. Banerjee, and E. Tutuc, Phys. Rev. B **83**, 161401 (2011).

[14] J. I. A. Li, T. Taniguchi, K. Watanabe, J. Hone, A. Levchenko, and C. R. Dean, Phys. Rev. Lett. **117**, 046802 (2016).

[15] M. Greenaway, E. Vdovin, A. Mishchenko, O. Makarovsky, A. Patanè, J. Wallbank, Y. Cao, A. Kretinin, M. Zhu, S. Morozov, V. I. Falko, K. Novoselov, A. Geim, T. Fromhold, and L. Eaves, Nat. Phys. **11**, 1057 (2015).

[16] L. Britnell, R. Gorbachev, A. Geim, L. Ponomarenko, A. Mishchenko, M. Greenaway, T. Fromhold, K. Novoselov, and L. Eaves, Nat. Commun. **4**, 1794 (2013).

[17] A. Mishchenko, J. Tu, Y. Cao, R. Gorbachev, J. Wallbank, M. Greenaway, V. Morozov, S. Morozov, M. Zhu, S. Wong, F. Withers, C. R. Woods, Y.-J. Kim, K. Watanabe, T. Taniguchi, E. E. Vdovin, O. Makarovsky, T. Fromhold, V. Fal'ko, A. Geim, L. Eaves, and K. Novoselov, Nat. Nanotechnol. **9**, 808 (2014).

[18] J. Gaskell, L. Eaves, K. S. Novoselov, A. Mishchenko, A. K. Geim, T. M. Fromhold, and M. T. Greenaway, Appl. Phys. Lett. **107**, 103105 (2015).

[19] R. M. Feenstra, D. Jena, and G. Gu, J. Appl. Phys. **111**, 043711 (2012).

[20] L. Brey, Phys. Rev. Applied **2**, 014003 (2014).

[21] F. T. Vasko, Phys. Rev. B **87**, 075424 (2013).

[22] E. E. Vdovin, A. Mishchenko, M. T. Greenaway, M. J. Zhu, D. Ghazaryan, A. Misra, Y. Cao, S. V. Morozov, O. Makarovsky, T. M. Fromhold, A. Patanè, G. J. Slotman, M. I. Katsnelson, A. K. Geim, K. S. Novoselov, and L. Eaves, Phys. Rev. Lett. **116**, 186603 (2016).

[23] B. Amorim, R. M. Ribeiro, and N. M. R. Peres, Phys. Rev. B **93**, 235403 (2016).

[24] D. Yadav, S. B. Tombet, T. Watanabe, S. Arnold, V. Ryzhii, and T. Otsuji, 2D Mater. **3**, 045009 (2016).

[25] J. Lambe and S. L. McCarthy, Phys. Rev. Lett. **37**, 923 (1976).

[26] B. Laks and D. L. Mills, Phys. Rev. B **20**, 4962 (1979).

[27] M. Parzefall, P. Bharadwaj, and L. Novotny, Antenna-coupled tunnel junctions, in *Quantum Plasmonics*, edited by S. I. Bozhevolnyi, L. Martin-Moreno, and F. Garcia-Vidal (Springer International Publishing, Cham, Switzerland, 2017), pp. 211–236.

[28] E. H. Hwang and S. D. Sarma, Phys. Rev. B **80**, 205405 (2009).

[29] T. Stauber and G. Gomez-Santos, Phys. Rev. B **85**, 075410 (2012).

[30] B. Sensale-Rodriguez, Appl. Phys. Lett. **103**, 123109 (2013).

[31] D. Svintsov, Z. Devizorova, T. Otsuji, and V. Ryzhii, Phys. Rev. B **94**, 115301 (2016).

[32] K. A. Guerrero-Becerra, A. Tomadin, and M. Polini, Phys. Rev. B **93**, 125417 (2016).
[33] V. Ryzhii, A. A. Dubinov, V. Y. Aleshkin, M. Ryzhii, and T. Otsuji, Appl. Phys. Lett. **103**, 163507 (2013).
[34] E. M. Belenov, P. N. Luskinovich, V. I. Romanenko, A. G. Sobolev, and A. V. Uskov, Sov. J. Quantum Electron. **17**, 1348 (1987).
[35] C. Zhang, M. L. F. Lerch, A. D. Martin, P. E. Simmonds, and L. Eaves, Phys. Rev. Lett. **72**, 3397 (1994).
[36] A. V. Uskov, J. B. Khurgin, I. E. Protsenko, I. V. Smetanin, and A. Bouhelier, Nanoscale **8**, 14573 (2016).
[37] C. B. Duke, M. J. Rice, and F. Steinrisser, Phys. Rev. **181**, 733 (1969).
[38] A. Bostwick, F. Speck, T. Seyller, K. Horn, M. Polini, R. Asgari, A. H. MacDonald, and E. Rotenberg, Science **328**, 999 (2010).
[39] M. B. Lundeberg, Y. Gao, A. Woessner, C. Tan, P. Alonso-González, K. Watanabe, T. Taniguchi, J. Hone, R. Hillenbrand, and F. H. Koppens, Nat. Mater. **16**, 204 (2017).
[40] F. J. G. de Abajo, Rev. Mod. Phys. **82**, 209 (2010).
[41] L. Zheng and S. D. Sarma, Phys. Rev. B **53**, 9964 (1996).
[42] E. H. Hwang and S. D. Sarma, Phys. Rev. B **75**, 205418 (2007).
[43] B. Wunsch, T. Stauber, F. Sols, and F. Guinea, New J. Phys. **8**, 318 (2006).
[44] L. Zheng and A. H. MacDonald, Phys. Rev. B **47**, 10619 (1993).
[45] R. Kazarinov and R. Suris, Sov. Phys. Semicond. **5**, 707 (1971).
[46] E. E. Salpeter, Phys. Rev. **120**, 1528 (1960).
[47] The interband tunneling is generally weaker than intraband due to the vanishing overlap of chiral wave functions $u^{+-}_{\mathbf{pp}'}$ in the limit of small momentum transfer.
[48] M. B. Lundeberg, Y. Gao, R. Asgari, C. Tan, B. Van Duppen, M. Autore, P. Alonso-González, A. Woessner, K. Watanabe, T. Taniguchi, R. Hillenbrand *et al.*, Science **357**, 187 (2017).
[49] V. Ryzhii, A. Satou, and T. Otsuji, J. Appl. Phys. **101**, 024509 (2007).
[50] L. Esaki, L. L. Chang, P. J. Stiles, D. F. O'Kane, and N. Wiser, Phys. Rev. **167**, 637 (1968).
[51] A. Principi, R. Asgari, and M. Polini, Solid State Commun. **151**, 1627 (2011).
[52] F. Rana, J. H. Strait, H. Wang, and C. Manolatou, Phys. Rev. B **84**, 045437 (2011).
[53] A. A. Dubinov, V. Y. Aleshkin, V. Mitin, T. Otsuji, and V. Ryzhii, J. Phys.: Condens. Matter **23**, 145302 (2011).
[54] The direct emission of photons is less probable due to the large spatial extent of photonic modes and small photonic density of states.
[55] M. Parzefall, P. Bharadwaj, A. Jain, T. Taniguchi, K. Watanabe, and L. Novotny, Nat. Nanotechnol. **10**, 1058 (2015).
[56] F. Stern, Phys. Rev. Lett. **18**, 546 (1967).
[57] K. Kempa, E. Gornik, K. Unterrainer, M. Kast, and G. Strasser, Phys. Rev. Lett. **86**, 2850 (2001).
[58] S. Gangadharaiah, A. M. Farid, and E. G. Mishchenko, Phys. Rev. Lett. **100**, 166802 (2008).
[59] L. Zheng and A. H. MacDonald, Phys. Rev. B **48**, 8203 (1993).
[60] E. H. Hwang, B. Y.-K. Hu, and S. D. Sarma, Phys. Rev. Lett. **99**, 226801 (2007).
[61] R. Z. Vitlina and A. V. Chaplik, JETP Lett. **78**, 651 (2003).
[62] W. Kwon, Y.-H. Kim, C.-L. Lee, M. Lee, H. C. Choi, T.-W. Lee, and S.-W. Rhee, Nano Lett. **14**, 1306 (2014).
[63] D. Ansell, I. Radko, Z. Han, F. Rodriguez, S. Bozhevolnyi, and A. Grigorenko, Nat. Commun. **6**, 8846 (2015).
[64] S. A. Khorasani, IEEE J. Quantum Electron. **50**, 307 (2014).
[65] A. Principi, M. Carrega, M. B. Lundeberg, A. Woessner, F. H. L. Koppens, G. Vignale, and M. Polini, Phys. Rev. B **90**, 165408 (2014).
[66] F. T. Vasko and V. Ryzhii, Phys. Rev. B **76**, 233404 (2007).
[67] S. Adam, E. Hwang, E. Rossi, and S. D. Sarma, Solid State Commun. **149**, 1072 (2009).