Emission and amplification of surface plasmons in resonant - tunneling van der Waals heterostructures

D. Svintsov¹, Zh. Devizorova²,³, T. Otsuji⁴, and V. Ryzhii⁴

¹Laboratory of Nano-optics and Plasmonics, Moscow Institute of Physics and Technology, Dolgoprudny 141700, Russia
²Department of Physical and Quantum Electronics, Moscow Institute of Physics and Technology, Dolgoprudny 141700, Russia
³Kotelnikov Institute of Radio Engineering and Electronics, Russian Academy of Science, Moscow, 125009 Russia and
⁴Research Institute of Electrical Communication, Tohoku University, Sendai 980-8577, Japan

We predict a new mechanism of surface plasmon amplification in graphene-insulator-graphene van der Waals heterostructures. The amplification occurs upon the stimulated interlayer electron tunneling accompanied by the emission of a coherent plasmon. The quantum-mechanical calculations of the non-local high-frequency tunnel conductivity show that a relative smallness of the tunneling exponent can be compensated by a strong resonance due to the enhanced tunneling between electron states with collinear momenta in the neighboring graphene layers. With the optimal selection of the barrier layer, the surface plasmon gain due to the inelastic tunneling can compensate or even exceed the loss due to both Drude and interband absorption. The tunneling emission of the surface plasmons is robust against a slight twist of the graphene layers and might explain the electroluminescence from the tunnel-coupled graphene layers observed in the recent experiments.

The ultrarelativistic nature of electrons in graphene gives rise to the uncommon properties of their collective excitations – surface plasmons [1, 3]. The deep subwavelength confinement [4], the unconventional density dependence of frequency [4, 5], and the absence of Landau damping [4] are probably the most well-known features of plasmons in graphene-based heterostructures. Among more sophisticated predictions there stand the existence of weakly damped transverse electric plasmons [6] and quasi-neutral electron-hole sound waves near the neutrality point [7, 8]. Some peculiar types of plasmons can be excited in the graphene p – n junctions [9], field-effect transistors [10, 11], optoelectronic modulators [12], and nanomechanical resonators [13] engaging for the improved device performance at the terahertz frequencies.

Unfortunately, the experimental studies of graphene plasmons are yet unable to confirm or refute many of these predictions. To achieve an extreme plasmon confinement, one has to sacrifice their propagation length. The latter is of the order of several micrometers at the infrared frequencies [14, 15] and is limited by the interband absorption in intrinsic samples and somewhat lower Drude absorption in the doped ones [16]. The experimentally reported quality factors of graphene plasmons reach only five for graphene on SiO₂ [14, 15], and 25 for graphene encapsulated in hexagonal boron nitride [17] at room temperature. In the latter case, the damping is due to the scattering by the intrinsic acoustic phonons [18] and can be suppressed only by lowering the temperature.

Instead of reducing the plasmon loss, it is possible to overcome the damping by introducing the gain medium which can replenish the energy being dissipated upon scattering. This idea has stimulated the re-examination of various ‘classical’ plasma instabilities in graphene, including the beam and resistive instabilities [19]. Dyakonov-Shur self-excitation [10, 11], and generation due to the negative differential conductance [19]. On the other hand, the plasmon gain can be provided by the photogenerated electrons and holes recombining with plasmon emission [20], which opens up the prospects of graphene-based spasers [21]. However, relatively fast nonradiative recombination in graphene [22, 23] hinders the practical implementation of those structures.

In this paper, we study the inelastic tunneling in graphene-dielectric-graphene van der Waals heterostructures accompanied by the emission of surface plasmons (SPs), and the possibility of the coherent SP amplification due to the tunneling gain. The inelastic plasmon-assisted tunneling was first observed in the late 70’s; it was shown to be responsible for the light emission from metal-insulator-metal tunneling diodes [24]. Afterwards, the tunneling excitation of SPs was demonstrated using the scanning probes above the metal surfaces [25, 26]. However, the possibility of the SP gain due to the tunneling to exceed the loss has been never considered realistic due to the smallness of the tunnel exponent, thus all experiments to date have reported only on the spontaneous SP emission. To achieve large tunneling gain, one needs some resonant feature to compensate the smallness of the tunnel exponent. As an example, such a resonance occurs in the quantum cascade lasers due to the alignment of energy subbands in the neighboring quantum wells [27–30]. In this paper, we show that the frequency dependence of plasmon tunneling gain in double graphene layer structures also exhibits a strong resonance due to the enhanced interaction between the Dirac electrons having collinear momenta [31]. In twisted graphene layers, the finite momentum of emitted plasmon can bridge the electron dispersions in the neighboring layers together [32], which makes the effect of plasmon tunneling emission ro-
bust against a slight layer twisting. We also discuss the recent experimental observations of the terahertz emission from the double graphene layer heterostructures and address the role of the inelastic-tunneling plasmon excitation in the observed spectra.

The double-graphene layer heterostructures shown schematically in Fig. 1 a support optical and acoustic surface plasmon modes having symmetric and antisymmetric distributions of electric potential. Their spectra are studied in detail in the absence of tunneling: the dispersion of symmetric (optical) mode is square-root, \( \omega_+ \approx v_0 \sqrt{3 \alpha_c q_0} \), where \( v_0 \) is the Fermi velocity in graphene, \( q_0 \) is the Fermi wave vector, \( \alpha_c = e^2/k \) is the coupling constant, while the dispersion of the antisymmetric (acoustic) mode is sound-like, \( \omega_- = s q \) (Fig. 1b). Its velocity approaches the Fermi velocity \( v_0 \) as the interlayer distance \( d \) decreases, but newer falls below it into the region of Landau damping. In realistic tunnel-coupled double layers, \( (d \approx 3 \text{ nm}, \epsilon_F \approx 0.1 \text{ eV}) \), \( s = 1.15 \epsilon_0 \). Just the acoustic mode has a nonzero average field strength between the two layers and thus can induce interlayer tunneling, a process in which an electron passes from one layer to another with emission of a coherent plasmon (Fig. 1 c). The tunneling can be included into the dispersion law of acoustic SPs by considering the tunnel current density \( J_{\perp q_{\omega}} \) as a generation-recombination term in the continuity equations for electrons in a single layer. In the linear response mode, this current is proportional to the interlayer potential difference, \( J_{\perp q_{\omega}} = G_{\omega} (\varphi_{t,q} - \varphi_{b,q}) \) where \( G_{\omega} \) is the tunnel conductance. In these notations, the dispersion law of the acoustic SP mode becomes

\[
-\frac{i\omega \kappa}{4\pi q} \left( 1 + \coth \frac{qd}{2} \right) + \sigma_{q_{\omega}} + \frac{2G_{\omega}}{q^2} = 0, \quad (1)
\]

where \( \sigma_{q_{\omega}} \) is the in-plane conductivity of graphene. Its real part, \( \text{Re} \sigma_{q_{\omega}} \), is positive and leads to the SP damping. At the same time, \( \text{Re} G_{q_{\omega}} \) can be negative, at least, at low energy of SP quanta \( \hbar \omega \). The negative tunnel conductance appears as the number of electrons with the energy \( E \) in top layer exceeds the number of electrons with energy \( E - \hbar \omega \) in the bottom layer; such carrier distribution can be viewed as a 'remote population inversion'. The sign of the quantity \( \text{Re} \left[ \sigma_{q_{\omega}} + \frac{2G_{q_{\omega}}}{q^2} \right] \) thus determines whether the SPs propagating along the double-layer are being damped or amplified.

An only missing ingredient to judge on the possibility of the net plasmon gain is the theory of non-local \( (q \neq 0) \) high-frequency tunnel conductivity of van der Waals structures; the latter, to the best of our knowledge, has been studied only in the DC limit. To construct this theory, we start with the tight-binding Hamiltonian of the tunnel coupled layers in the presence of the propagating plasmon (we set \( \hbar \equiv 1 \)):

\[
\hat{H} = \left( H_{G_G} + \hat{T}^* \hat{H}_G - \hat{T} \right) + \hat{H}_{\text{int}} \equiv \hat{H}_0 + \hat{H}_{\text{int}}. \quad (2)
\]

Here, \( \hat{H}_{G_G} = v_0 \sigma \hat{p} \pm i \Delta/2 \) describe separate graphene layers, \( \Delta \) is the energy spacing between the Dirac points, \( \hat{p} \) is the in-plane momentum operator, \( \hat{I} \) is the identity matrix, and \( \hat{T} \) is the tunneling matrix. For the
sake of analytical traceability, we choose $\hat{T}$ in its simplest form which is applicable for AA-stacking of aligned graphene layers \[33, 40\], $\hat{T} = \Omega \hat{I}$, where $\Omega$ is the ‘tunneling frequency’. The interaction part, $\hat{H}_{\text{int}}$, describes the electron-plasmon coupling; its matrix elements are calculated as the overlap of the eigen functions of $\hat{H}_0$ with the electron potential energy in the field of SP $v_\varphi(x, z, t) = e\varphi_0 s(z)e^{i(qz-\omega t)}$, where $\varphi_0$ is the amplitude of potential on the top layer and $s(z)$ is the dimensionless ‘shape function’ (see Fig. 1 b).

A relatively simple form of the $\hat{T}$-matrix allows us to treat the tunneling non-perturbatively \[41\]. The good quantum numbers of the eigen states of $\hat{H}_0$ are the in-plane electron momentum $\mathbf{p}$, the band index $s = \pm 1$ ($s = +1$ for the conduction and $s = -1$ for the valence band), and an extra index $l = \pm 1$ governing the $z$-localization of electron wave function. At large bias voltage, $l = +1$ and $l = -1$ correspond to the states localized on the top and bottom layers, respectively, while at small bias $l = +1$ and $l = -1$ describe the anti-symmetric and symmetric states of the tunnel-coupled quantum wells. The eigen energies of these states are \[3\]

$$\varepsilon_p^s = s v_0 p + l \sqrt{\Omega^2 + \Delta^2 / 4}. \tag{3}$$

The evaluation of tunnel current density is based on the relation $J_{\perp, q, \omega} = g T_0 \left[ J_{\perp, \alpha \beta} \rho^{(1)}_{\beta \alpha} \hat{p}_{\mathbf{p} + \mathbf{q}, \mathbf{p}'} \right]$, where $\rho^{(1)}$ is the electron density matrix calculated up to the first order in electron-plasmon interaction, $g = 4$ is the spin-valley degeneracy factor, and the indices $\alpha = \{p, l, s\}$, $\beta = \{p', l', s'\}$ run over all quantum numbers. The explicit form of the tunnel current operator is found from the ‘equation of motion’, $\dot{\hat{J}}_\perp = d\hat{Q}_t / dt = i [\hat{H}, \hat{Q}_t]$, where $\hat{Q}_t$ is the operator of electron charge in the top layer. The density matrix $\rho^{(1)}$ is found from the von Neumann equation

$$i \frac{\partial \rho^{(1)}}{\partial t} = \left[ \hat{H}_0, \rho^{(1)} \right] + \left[ \hat{H}_{\text{int}}, \rho^{(0)} \right], \tag{4}$$

being interested in the linear response, we commute only the zeroth-order density matrix $\rho^{(0)}$ with $\hat{H}_{\text{int}}$. Considering the harmonic time dependence, Eq. (4) is readily solved in the diagonal basis of $\hat{H}_0$ leading to a closed-form expression for the frequency-dependent non-local tunnel conductance

$$G_{q, \omega} = i g e^2 \Omega \sum_{\mathbf{p}, \mathbf{p}', \mathbf{q}} u_{\mathbf{p} \mathbf{p}'}^{s s'} \frac{v_{\mathbf{p} \mathbf{p}'}^{s s'} s_{\mathbf{p} \mathbf{p}'}^{\omega} \left[ f_{\mathbf{p}'}^{s'} - f_{\mathbf{p}'}^{s''} \right]}{\omega + i \delta - \varepsilon_{\mathbf{p} \mathbf{p}'}^{s s'} - \varepsilon_{\mathbf{p} \mathbf{p}'}^{s'' s''}}. \tag{5}$$

Here, $s_{\mathbf{p} \mathbf{p}'}^{\omega}$ are the overlap integrals between the SP field profile and the $z$-components of the electron wave functions, $s_{\mathbf{p} \mathbf{p}'}^{\omega} = \int_{-\infty}^{\infty} dz \hat{\Psi}_{\mathbf{p}}^{s} (z) \hat{\Psi}_{\mathbf{p}'}^{s'} (z)$ and $u_{\mathbf{p} \mathbf{p}'}^{s s'}$ are the overlap factors between the chiral envelope wave functions in graphene, $u_{\mathbf{p} \mathbf{p}'}^{s s'} = \left[ 1 + s s' \cos \theta_{\mathbf{p} \mathbf{p}'} \right] / 2$, and $f_{\mathbf{p}'}^{s''}$ are the occupation numbers of the eigen states of given by the Fermi functions shifted by the applied bias $eV$ in the energy scale.

At this point, it is noteworthy to mention the similarity of Eq. (5) and the expressions for the graphene polarizability \[3\]. The latter diverge on the ‘Dirac cone’ as $[\omega^2 - q^2 v_0^2]^{-1/2}$, and a similar singularity appears in the tunnel conductance \[\text{Eq. (5)}\], except the frequency $\omega$ should be replaced with $\omega - \sqrt{\Delta^2 + 4\Omega^2}$. This singularity is simply elucidated after extracting the real part of Eq. (5) and passing to the elliptic coordinates, which leads us to the following expression

$$\text{Re} G_{q, \omega} = - \frac{e^2}{4 \pi \hbar} \frac{q^2 s_{\perp \omega} \Omega \omega}{(q v_0)^2 - (e V_{12} - \omega)^2}, \tag{6}$$

where we have introduced the effective energy shift between the Dirac points $e V_{12} = \sqrt{\Delta^2 + 4\Omega^2}$, and a dimensionless non-singular function $\mathcal{I}$ appearing as a result of the Fermi distribution integration,

$$\mathcal{I} = \sinh \left( \frac{eV - \omega}{T} \right) \int_1^\infty \frac{\sqrt{t^2 - 1}}{\cosh \left( \frac{q v_0}{T} t \right) + \cosh \left( \frac{e v_{12} - \omega}{T} \right)}. \tag{7}$$

The real part of the tunnel conductance is negative provided $\omega < eV$, which implies that the tunneling electrons rather loose their energy by emitting a plasmon than get the energy by plasmon absorption. At the same time, the tunnel transitions from top layer to the bottom one accompanied by SP absorption are prohibited by the energy conservation law as far as the SP velocity exceeds the Fermi velocity. Still, just like in most tunnel phenomena, the negative conductivity due to the tunneling is proportional to a small exponent $e^{-2\kappa d}$ ($\kappa$ is the decay constant of the electron wave function) that eventually enters $s_{\perp \omega}$ and $\Omega$. To enhance the negative conductance, the materials with small effective (tunneling) mass $m^*$ and small band offset $U_b$ with respect to graphene have to be used. Boron nitride ($U_b = 1.5$ eV, $m^* = 0.5 m_0$) is not the best candidate for the realization of the net SP gain, while chalcogenides of molybdenum MoS$_2$ ($U_b = 0.29$ eV, $m^* = 0.43 m_0$) and tungsten WS$_2$ ($U_b = 0.4$ eV, $m^* = 0.28 m_0$) \[42\] demonstrate significantly stronger tunneling.

The smallness of tunnel current is to a considerable extent compensated by a singularity in plasmon gain, which occurs provided $e V_{12} = \omega + q v_0 = \omega (1 + v_0 / s)$. The structure of the $p$-integral in Eq. (5) shows that the singularity comes from the tunneling between electron states with collinear momenta in the neighboring layers. Once the carrier dispersion is linear, these states have the same velocity and thus interact for an infinitely long time \[31\]. The carrier scattering naturally destroys this coherence making the singular tunnel conductance finite. To account for this effect quantitatively, we add the imaginary self-energy corrections $\text{Im} \Sigma (p, E)$ to the quasi-particle energies in \[3\]. These corrections appear
as a result of electron-phonon scattering, the latter also leading to the finite Drude absorption. For simplicity, we take \( \text{Im} \Sigma(p, E) \) to be energy- and momentum independent and equal to its value at the Fermi-surface \( \text{Im} \Sigma(p_F, \epsilon_F) = \nu \),

\[
\nu = \frac{\epsilon_F TD^2}{4\rho v_0^2},
\]

(8)

where \( D \approx 30 \text{ eV} \) is the deformation potential of graphene \( [43] \), \( \rho = 7.6 \times 10^{-7} \text{ kg/m}^2 \) is its mass density, and \( \epsilon \approx 0.02v_0 \) is the sound velocity \( [44] \).

To judge on the possibility of the net plasmon gain, one has also to evaluate the frequency-dependent non-local in-plane conductivity, \( \sigma_{\text{q}, \omega} \). In the absence of electron collisions, one can readily apply the Kubo linear response theory and obtain the well-known result including interband and intraband contributions \( [43] \). The real part of interband optical conductivity \( (q = 0) \) is universal at high frequencies, \( \text{Re} \sigma_{\text{inter}} = e^2/4\hbar \). However, at finite wave vector – which is the case of plasmons – the frequency dependence of conductivity becomes more complicated,

\[
\text{Re} \sigma_{\text{inter}}(q, \omega) \approx \frac{e^2}{4\hbar} \frac{\omega}{\omega^2 - q^2v_0^2} [f_0(-\omega/2) - f_0(\omega/2)].
\]

(9)

The neglect of the spatial dispersion results in an underestimate of \( \text{Re} \sigma_{\text{inter}} \) and, hence, plasmon damping. This is especially crucial for the acoustic modes which velocity just slightly exceeds the Fermi velocity and the respective interband conductivity is in a dangerous vicinity of the singularity at \( \omega = qv_0 \).

Considering the intraband (Drude) conductivity, a special care should be taken to account correctly both for the spatial dispersion and finite carrier relaxation time \( [10] \). Apparently, the effects of carrier collisions cannot be included by a simple replacement \( \omega \to \omega + i\nu \) in the Kubo-like expression for the collisionless conductivity \( [40] \). Even in the quasi-classical formalism of Boltzmann equation (which is justified in the case of interest, \( q \ll q_F \), \( \omega \ll \epsilon_F \)), the collision integral in the \( \tau \)-approximation violates the particle conservation. To avoid those difficulties, we have evaluated the conductivity by solving the kinetic equation with the particle-conserving Bhatnagar-Gross-Krook collision integral \( [47] \) in the right-hand side. The respective intraband dynamic conductivity reads

\[
\sigma_{\text{intra}}(q, \omega) = \frac{ie^2 \bar{\epsilon}_F}{\pi^2 \omega} J_2 \left( \frac{2\pi \nu qv_0}{\omega} \right) \frac{J_1 \left( \frac{2\pi \nu qv_0}{\omega} \right)}{1 - \frac{2\pi \nu qv_0}{\omega} J_1 \left( \frac{2\pi \nu qv_0}{\omega} \right)},
\]

(10)

where \( J_n(a, b) = \int_0^{2\pi} d\theta \cos^n \theta/[1 - a \cos \theta + ib] \), \( \bar{\epsilon}_F = T \ln(1 + e^{\epsilon_F/T}) \), and \( \nu \) is the collision frequency limited by the electron-phonon scattering \( [17] \) given essentially by Eq. (8). At zero wave vector, \( J_1 = 0 \), and we restore the Boltzmann conductivity \( \sigma_{\text{intra}}(0, \omega) = ie^2 \bar{\epsilon}_F/\pi(\omega + i\nu) \), while for nonzero \( q \) the real part of intraband conductivity typically appears to be larger than its zero-\( q \) value.

This difference between no-local and local conductivities agrees with the experimentally observed distinction between the plasmon lifetime and the Boltzmann relaxation time \( \nu^{-1} [17] \).

With these prerequisites, we are able to compare the plasmon gain due to the tunneling and loss due to the Drude and interband absorption. Namely, if in some frequency range the effective conductivity \( \sigma_{\text{eff}} = \text{Re} [\sigma_{\text{inter}} + \sigma_{\text{intra}} - 2G/q^2] \) is negative, then the plasmon gain exceeds loss. The self-excitation of plasmons with frequencies satisfying the net gain condition is possible. Figure 2 shows the wave vector- and frequency dependence of the effective conductivity at the room and nitrogen temperatures with the region under the green plane corresponding to the negative values. The two sets of singularities are clearly present in the conductivity spectra: the absorption singularity at \( \omega = qv_0 \) comes from the in-plane conductivity, and the gain singularity at \( \omega = eV_{12} \) comes from the enhanced tunneling between the collinear

![Figure 2](image-url)

FIG. 2. Net "effective conductivity" as it appears in the SP dispersion law, \( \text{Re}[\sigma_{\text{q}, \omega} + 2G_{\text{q}, \omega}/q^2] \) (in the units of \( e^2/h \)) vs. frequency and wave vector at two different temperatures: \( T = 77 \text{ K} \) (top) and \( T = 300 \text{ K} \) (bottom). The barrier layer is 2.5 nm WS₂. The singularities of the interlayer tunnel conductance, \( \hbar \omega = eV_{12} \pm qv_0 \) are shown with blue lines, the dispersion of acoustic SP is shown with red line. In the region filled with green, the net effective conductivity is negative.
states. As the wave vector approaches zero, the effective conductance grows indefinitely due to the presence of $q^2$ in the denominator, this will not, however, lead to the infinite plasmon gain as the acoustic SP spectra develop a gap $\Delta \omega \simeq \Omega$ at small wave vectors \cite{30}. On the other hand, there is a wide range of frequencies $\omega \gtrsim \Omega$ where the dispersion of acoustic SPs in not strongly affected by tunneling, but the real part of net effective conductivity is negative, which justifies the possibility of gain.

The spectra of in-plane conductivity $\text{Re}[\sigma_{\text{inter}} + \sigma_{\text{intra}}]$ and effective tunnel conductance $-2\text{Re}[G/q^2]$ are compared in Fig. 3 for the wave vectors satisfying the acoustic plasmon dispersion $\omega = sq$. At room temperature (upper panel) the self-excitation is possible in a narrow ($\sim 15 - 20 \text{ meV}$) vicinity of the tunneling resonance frequency, while at higher frequencies the interband absorption surpasses the gain. At lower temperature of 77 K (lower panel) the range of frequencies corresponding to the net SP gain is significantly broader, while the peak tunnel conductance is higher. The broadening of the gain region is due to efficient low-temperature Pauli blocking of the interband transitions with energy below the double Fermi energy. The sharpening of the resonant peak is attributed to the phonon freezing-out at low temperatures, which increases the quasiparticle lifetime and resonant gain. It is also noteworthy that the peak in the negative tunnel conductance occurring at $\omega_{\text{res}} = \epsilon V_{l2}/(1 + v_0/s) \approx \epsilon V_{l2}/2$ falls into the ‘transparency window’ of graphene where the Drude absorption is low ($\omega_{\text{res}} \gg \nu$) but the interband absorption is still blocked ($\omega_{\text{res}} < 2\epsilon_F$).

The efficiency of SP excitation can be further improved in the gated double-graphene layer structures, where the energy shift between Dirac points $\Delta$ and the carrier density can be controlled independently. At the fixed carrier density, the plasmon gain increases with reduction in $\Delta$ as the energies of the $l = +1$ and $l = -1$ states get closer in energy scale resulting in enhanced tunnel coupling.

In realistic graphene-based tunnel structures, the adjacent layers are always slightly misaligned in real space, which results in the misalignment of Dirac cones in the $k$-space. The finiteness of SP wave vector can, to some extent, compensate this misalignment. If the Dirac cones in adjacent layers are separated by a wave vector $q_M$, the tunneling resonance will be retained, though the position of the resonant peak will depend on the direction of plasmon propagation $q$. The resonant condition in this case is easily shown to be

$$|q + q_M|^2 v_0^2 = (\epsilon V_{l2} - \omega)^2,$$

and the expressions for the real part of tunnel conductance are obtained from Eq. (6) by a simple replacement $g \rightarrow |q + q_M|$. The case of the plasmon-assisted resonant tunneling is thus radically different from the photon-assisted tunneling \cite{30}, where even a slight misalignment breaks the resonance due to the negligible photon momentum. The robustness of SP emission against slight misalignment correlates with the broadness of gain (and emission) spectra extending above the resonant frequency. At the same time, the spectra of phonons emitted upon tunneling are predicted to be Lorentzian their width being limited by the carrier relaxation rate. For these reason, the inelastic tunneling accompanied by the emission of surface plasmons rather than photons may largely contribute to the spontaneous terahertz emission from the double graphene layer structures observed in \cite{32}. The mechanism of electromagnetic radiation in case of plasmon excitation could be either the dipole radiation of the whole heterostructure, or the radiative decay of plasmons scattered by the edges.

In conclusion, we have theoretically studied the frequency-dependent non-local tunnel conductivity in the biased graphene layers. In a wide range of frequencies, the real part of the tunnel conductivity is negative due
to the dominance of inelastic tunneling with surface plasmon emission over the absorption. Moreover, at frequencies satisfying the condition \( \hbar \omega = e V_{12} + q v_0 \), where \( V_{12} \) is the interlayer potential difference and \( q \) is the plasmon wave vector, the absolute value of negative tunnel conductivity is resonantly large, which is a manifestation of enhanced tunneling between electron states with collinear momenta in the neighboring layers. A detailed comparison of the surface plasmon loss due to the interband and Drude absorption and gain due to the inelastic tunneling shows the possibility of the net gain in sufficiently thin tunnel structures.

The work of DS was supported by the grant # 14-07-31315 of the Russian Foundation of Basic Research. The work at RIEC was supported by the Japan Society for Promotion of Science (Grant-in-Aid for Specially Promoted Research No. 23000008).

APPENDIX

Evaluation of the in-plane conductivity

The expression for the interband part of conductivity is readily obtained from the Kubo theory [45]

\[
\sigma_{\text{inter}}(q, \omega) = \frac{2i e^2 \omega}{\pi^2} \sum_{p, s \neq s'} \left[ v_{ps'}^+ - v_{ps}^+ \right] \left( f_{p'}^s - f_p^s \right) \left[ \omega + i \delta - (\epsilon_{p'}^s - \epsilon_p^s) \right]. \tag{A1}
\]

Here \( v_{ps'}^+ \) is the matrix element of velocity operator in graphene \( \vec{v} = v_0 \sigma \), \( p_\pm = p \pm q/2 \), \( f_p^s \) is the electron distribution functions in the valence \((s = -1)\) and the conduction \((s = +1)\) band, and \( \epsilon_p^s \) is the dispersion law in the \( s \)-th band. Known the eigen functions of graphene Hamiltonian \( \hat{H}_G = v_0 \sigma \vec{r} \cdot \vec{p} \),

\[
| \vec{p} \rangle = \frac{1}{\sqrt{2}} \left( e^{-i \theta_p^+/2} \right) \left( \begin{array}{c} p \cos \theta_p^+ \sin \theta_p^- \sin \theta_p^+ \cos \theta_p^- \end{array} \right), \tag{A2}
\]

one readily finds \( \langle \vec{p}_- | \vec{v} | \vec{p}_+ \rangle = iv_0 \sin (\theta_p^+ + \theta_p^-)/2 \). The subsequent calculations are conveniently performed in the elliptic coordinates

\[
p = \frac{q}{2} \left( \cos u \cos v, \sin u \sin v \right). \tag{A3}
\]

In these coordinates \( |p_\pm \rangle = (q/2) | \cos u \pm \cos v \rangle \), \( |p_- \rangle |p_- \rangle \rangle = (q v_0)^2 \cos^2 u \sin^2 v dudv \). The real part of conductivity is readily extracted with Sokhotski theorem, which leads us to

\[
\text{Re} \sigma_{\text{inter}}(q, \omega) \approx \frac{e^2}{4 \hbar} \sqrt{\omega^2 - q^2 v_0^2} \int dv \sin^2 v \times \left\{ f_0 \left[ \frac{-\omega}{2} + \frac{q v_0}{2} \cos v \right] - f_0 \left[ \frac{\omega}{2} + \frac{q v_0}{2} \cos v \right] \right\}. \tag{A4}
\]

To proceed further, we note that in the domain of interest \( \omega > q v_0 \) one always has \( q v_0 \cos v < \omega \). As a simplest approximation, one can set \( q = 0 \) in the arguments of distribution function and obtain

\[
\text{Re} \sigma_{\text{inter}}(q, \omega) \approx \frac{e^2}{4 \hbar} \frac{\omega}{\sqrt{\omega^2 - q^2 v_0^2}} \left[ f_0 (-\omega/2) - f_0 (\omega/2) \right]. \tag{A5}
\]

The agreement between approximate analytical expression \( \text{(A5)} \) and exact integral relation \( \text{(A4)} \) is fair as far as \( q v_0 \) is not too close to \( \omega \). A better agreement can be obtained after the change of variable \( \cos v = t \) in Eq. \( \text{(A4)} \) by noting that \( \sin v = \sqrt{1-t^2} \) varies rapidly over the integration domain \( t \in [-1; 1] \), while the remainder \( g(t) = f_0 (-\omega - q t)/2 - f_0 (\omega + q t)/2 \) varies slowly. Thus, one can make the following approximation

\[
\int_{-1}^{1} g(t) \sqrt{1-t^2} dt \approx \frac{1}{2} \int_{-1}^{1} \sqrt{1-t^2} dt \int_{-1}^{1} g(t) dt. \tag{A6}
\]

The following approximation for the conductivity holds

\[
\text{Re} \sigma_{\text{inter}}(q, \omega) \approx \frac{e^2}{4 \hbar} \frac{T}{q v_0} \frac{\omega}{\sqrt{\omega^2 - q^2 v_0^2}} \times \ln \left( \frac{\cosh \frac{\omega}{2} + \cosh \frac{q v_0}{2}}{\cosh \frac{\omega}{2} + \cosh \frac{q v_0}{2}} \right). \tag{A7}
\]

Figure \( \text{A1} \) shows the calculated interband conductivities using the exact \( \text{(A1)} \) and approximate expressions \( \text{(A5)} \) and \( \text{(A7)} \) as well as the local \( (q = 0) \) approximation. Clearly, the neglect of spatial dispersion in the case of...
of acoustic SPs with velocity slightly exceeding the Fermi velocity results in an underestimation of the damping.

We now pass to the in-plane conductivity associated with the intraband transitions, \( \text{Re} \sigma_{\text{intra}}(q, \omega) \). We shall restrict ourselves to the classical description of the intraband electron motion which is justified at frequencies \( \hbar \omega \ll \varepsilon_F, q \ll q_F \). The effect of plasmon tunneling gain occurs in this frequency domain, otherwise, the interband damping of SPs takes place [see Eq. (9)]. To account for the carrier relaxation in the non-local \((q \neq 0)\) case we adopt the formalism of kinetic equation with the particle-conserving Bhatnagar-Gross-Krook collision integral \([47]\) in the right-hand side,

\[
-i \omega \delta f_0(p) + i q v \delta f_0(p) + i e q v \delta \varphi_0 \frac{\partial f_0}{\partial \varphi} = - \nu \left[ \delta f_0(p) + \frac{d \varepsilon_F}{d n} \frac{\partial f_0}{\partial \varepsilon} \delta n_\varphi \right]. \quad (A8)
\]

Here \( \delta f_0(p) \) is the sought-for field-dependent correction to the equilibrium electron distribution function \( f_0 \), \( \delta n_\varphi \) is the respective correction to the electron density, \( v = v_0 p/p \) is the quasi-particle velocity, and \( \nu \) is the electron collision frequency which is assumed to be energy-independent. The current density, associated with the distribution function \( \delta f_0(p) \) reads:

\[
\delta j_q = - e q \sum_p \frac{d f_0}{d \varepsilon} \frac{i e q v \delta \varphi_0 - i v d \varepsilon}{\omega + i \nu - q v} \delta n_\varphi. \quad (A9)
\]

Recalling the relation between small-signal variations of density and current, \( \omega \delta n_\varphi = q \delta j_q \), and evaluating the integrals in Eq. (A9), we find the conductivity given by Eq. (10).

Spectra of plasmons coupled to the double graphene layer heterostructures

The plasmon spectra are obtained by a self-consistent solution of the Poisson's equation

\[
-q^2 \delta \varphi_0(z) + \frac{\partial^2 \delta \varphi_0(z)}{\partial z^2} = - \frac{4 \pi}{\kappa} \left[ \rho_{\text{el}, \varphi}(z - d/2) + \rho_{\text{el}, \varphi}(z + d/2) \right], \quad (A10)
\]

the continuity equations

\[
-i \omega \rho_\varphi - q \delta j_q = 0, \quad (A11)
\]

and the linear-response relation between current density and electric field, \( \delta j_q = \sigma_{q, \omega} E_q \). Here \( q \) is the two-dimensional plasmon wave vector, \( d \) is the distance between layers, \( \kappa \) is the background dielectric permittivity, \( \rho_{\text{el}, \varphi} \) and \( \rho_{\text{el}, \varphi} \) are the small-signal variations of charge density in the top and bottom layers, respectively. In the absence of built-in voltage, due to the electron-hole symmetry, the charge densities in the layers are equal in modulus as opposite in sign, moreover, the layer conductivities are equal. This allows us to seek for the solutions of Eq. (A10) being symmetric and anti-symmetric with respect to the electric potential. A straightforward calculation brings us to the following dispersions \([55]\)

\[
-\frac{i \omega \kappa}{4 \pi} \left( 1 + \coth \frac{q d}{2} \right) + q \sigma_{q, \omega} = 0 \quad (A12)
\]

for the antisymmetric (acoustic) mode, and

\[
-\frac{i \omega \kappa}{4 \pi} \left( 1 + \tanh \frac{q d}{2} \right) + q \sigma_{q, \omega} = 0 \quad (A13)
\]

for the symmetric (optical mode). Being interested in the long-wavelength limit, \( q d/2 \ll 1 \), we perform the expansions \( 1 + \coth q d/2 \approx 2/q d, 1 + \tanh q d/2 \approx 1 \). In the same limit, the conductivity is essentially classical, moreover, the interband transitions do not affect the low-energy part of the spectra. With these assumptions, we use the following (collisionless) approximation for the conductivity

\[
\sigma_{q, \omega} = i g^2 \frac{\varepsilon_F}{\pi} \frac{\omega}{q^2 v_F^2} \left[ \frac{\omega}{\sqrt{\omega^2 - q^2 v_F^2}} - 1 \right], \quad (A14)
\]

which follows readily from Eq. (10) after setting \( \nu = 0 \). Equation (A12) admits an analytical solution, which is a sound-like dispersion

\[
\omega_- = v_0 \sqrt{1 + 4 \alpha_q q d/\sqrt{1 + 8 \alpha_q q d}}. \quad (A15)
\]

Here, we have introduced the Fermi wave vector \( q_F = \varepsilon_F/v_0 \). The velocity of the acoustic mode always exceeds the Fermi velocity thus preventing the Landau damping. The dispersion equation for the optical mode \( \omega_+(q) \) is cubic, however, in the long-wavelength limit the spatial dispersion of conductivity can be neglected as the phase velocity of this mode significantly exceeds the Fermi velocity. The approximate relation for \( \omega_+(q) \) has the following form

\[
\omega_+ \approx v_0 \sqrt{4 \alpha_q q d}. \quad (A16)
\]

To account for the damping of the propagating plasmon one can express the wave vector \( q \) from Eqs. (A12) and (A13) and consider the real part of conductivity along with the imaginary one. The results of such calculations are shown in Fig. (A2) along with the frequency dependencies of the net conductivity. The acoustic mode is typically damped more heavily than optical one the two effects being responsible for that fact. First of all, the velocity of optical mode is higher than that of acoustic one, hence, at a fixed scattering time, the free path of optical mode is higher. The second reason is the increase
where $\phi$ is the electric potential on the top layer. The shape function has the following form

$$s(z) = \begin{cases} e^{-k(z+d/2)}, & z < -d/2, \\ \sinh(qz/2)/\sinh(qd/2), & |z| < d/2, \\ -e^{-k(z-d/2)}, & z > d/2. \end{cases}$$

(A18)

Estimate of the effective tight-binding parameters

The tight-binding Hamiltonian of the tunnel-coupled graphene layers in the absence of the propagating plasmon Hamiltonian $\hat{H}_0$ in Eq. (2) constitutes the blocks describing isolated graphene layers $\hat{H}_{G\pm}$ and the block describing tunnel hopping $\hat{T}$. This Hamiltonian acts on a four-component wave function

$$\psi = \{\psi_{A,t}, \psi_{B,t}, \psi_{A,b}, \psi_{B,b}\}^T$$

(A19)

which components are the probability amplitudes of finding an electron on a definite layer and on the definite lattice cites, $A$ or $B$. Such description of electron states is common for graphene bilayer; moreover, the elements of tunneling matrix have been evaluated for different twist angles between single layers constituting a bilayer. In more comprehensive theories, the $T$-matrix is affected by the band structure of dielectric layer. Here, for the sake of analytical traceability, we choose the tunneling matrix in its simplest form which is applicable to the AA-stacked perfectly aligned graphene bilayer, $T = \Omega I$, where $\Omega$ can be interpreted as the tunnel hopping frequency.

To estimate its value, we switch for a while from the tight binding to the continuum description of electron states in the $z$-direction. We model each graphene layer with a delta-well $U(z) = A\delta(z - z_t)$, where $A$ is the potential strength chosen to provide a correct value of electron work function $U_b$ from graphene to the background dielectric,

$$A = 2\sqrt{\frac{\hbar^2 U_b}{2m^*}}.$$  

(A20)

where $m^*$ is the effective electron mass in the dielectric. The effective Schrodinger equation in the presence of voltage bias $\Delta/e$ between graphene layers takes on the following form

$$-\frac{\hbar^2}{2m^*} \frac{\partial^2 \Psi(z)}{\partial z^2} + 2\sqrt{\frac{\hbar^2 U_b}{2m^*}} \left[\delta(z - d/2) + \delta(z + d/2)\right] \Psi(z) + U_F(z) \Psi(z) = E \Psi(z),$$

(A21)

where $U_F$ is the potential energy created by the applied field

$$U_F(z) = \frac{\Delta}{2} \begin{cases} 1, & z < -d/2, \\ 2z/d, & |z| < d/2, \\ -1, & z > d/2. \end{cases}$$

(A22)

The solutions of effective Schrodinger equation represent decaying exponents at $|z| > d/2$, and a linear combination of Airy functions in the middle region $|z| < d/2$

$$\Psi_M(z) = C\text{Ai}(-z/a + \varepsilon) + DB\text{Bi}(-z/a + \varepsilon),$$

(A23)

where $\varepsilon = 2m^*|E|a^2/\hbar^2$ is the dimensionless energy and $a = (\hbar^2d/2m^*\Delta)^{1/3}$ is the effective length in the electric field. A straightforward matching of the wave functions at the graphene layers yields the dispersion equation
where \( k_b = \sqrt{2m^*U_b/h^2} \) is the decay constant of the bound state wave function in a single delta-well, \( k_1 = \sqrt{2m^*(E + \Delta/2)/h^2}, \ k_2 = \sqrt{2m^*(E - \Delta/2)/h^2} \). Equation (A24) yields two energy levels \( E_l (l = \pm 1) \) which cannot be expressed in a closed form via the parameters of the well. However, the dependence of \( E_l \) on the energy separation between layers \( \Delta \) is accurately described by

\[
E_l(\Delta) = -U_b + \frac{l}{2} \sqrt{(E_{+1,\Delta=0} - E_{-1,\Delta=0})^2 + \Delta^2}. \tag{A25}
\]

This allows us to estimate the tunnel coupling \( \Omega \) as half the energy splitting of states in double graphene layer well in the absence of applied bias

\[
\Omega = \frac{1}{2} [E_{+1,\Delta=0} - E_{-1,\Delta=0}]. \tag{A26}
\]

The \( l \)-index governs the \( z \)-localization of electron in a biased double quantum well. At large bias \( \Delta \gg \Omega \), the delta-wells interact weakly, thus \( l = +1 \) corresponds to the state localized almost completely in the top layer and \( l = -1 \) corresponds to the electron in the bottom layer. At small bias \( \Delta \approx \Omega \) the state \( l = +1 \) is antisymmetric and \( l = -1 \) is symmetric.

Electron-plasmon interaction and solution of the von Neumann equation

The presence of plasmon propagating along the double graphene layer results in an additional potential energy of electron

\[
H_{\text{int}}(r, t) = \epsilon \varphi(z) e^{i(qz - \omega t)}, \tag{A27}
\]

where we assume the direction of plasmon propagation to be along the \( x \)-axis, and the dependence of potential on the \( z \)-coordinate is given by Eqs. (17) and (18). The additional terms in Hamiltonian due to the vector-potential are negligible as far as the speed of light substantially exceeds the plasmon velocity.

With our choice of the tight-binding basis functions (19) as those localized on a definite layer and on a definite lattice cite, we shall require 16 matrix elements of the potential energy (28) connecting those basis states. However, it is more convenient to work out the matrix elements of (28) connecting the eigenstates of Hamiltonian (2). The good quantum numbers of these states are the in-plane momentum \( p \), the band index \( s = \pm 1 (+1 \) for the conduction band and \( -1 \) for the valence band) and the \( l \)-index discussed above. The respective matrix elements are

\[
\langle p, s, l | \hat{H}_{\text{int}} | p' s' l' \rangle = \delta_{p,p'} \delta_{s,s'} \epsilon_{p} \int_{-\infty}^{\infty} \Psi^*_l(z) s(z) \Psi_{l'}(z). \tag{A29}
\]

Having obtained the matrix elements of electron-plasmon interaction, we pass to the solution of the von Neumann equation (4). Being interested in the linear response of electron system to the plasmon field, we decompose the density matrix as \( \hat{\rho} = \hat{\rho}^{(0)} + \hat{\rho}^{(1)} \), where \( \hat{\rho}^{(1)} \) is linear in the electron-plasmon interaction \( \hat{H}_{\text{int}} \). The calculations are conveniently performed in the basis of the eigenstates of \( \hat{H}_0 \) including the effects of tunneling nonperturbatively. In this basis, \( [\hat{H}_0, \hat{\rho}^{(1)}]_{\alpha\beta} = (\varepsilon_\alpha - \varepsilon_\beta) \rho^{(1)}_{\alpha\beta} \), and thus one readily writes down the solution for the density matrix

\[
\rho^{(1)}_{\alpha\beta} = \frac{[\hat{H}_{\text{int}}, \hat{\rho}^{(0)}]_{\alpha\beta}}{\omega + i \delta - (\varepsilon_{\alpha} - \varepsilon_{\beta})}. \tag{A30}
\]

The first-order correction is now expressed through the density matrix in the absence of plasmon field \( \hat{\rho}^{(0)} \). A particular choice of \( \hat{\rho}^{(0)} \) requires the solution of kinetic equation in the voltage-biased tunnel-coupled layers, however, in several limiting cases the situation is greatly simplified (28). If the tunneling rate \( \Omega \) is slower than the electron energy relaxation rate \( \nu_e \) (e.g., due to phonons and carrier-carrier scattering), the quasi-equilibrium distribution function is established in each individual layer. In this situation, \( \rho^{(0)} \) is diagonal in the basis formed by the wave functions localized on top and bottom layers its elements being the respective Fermi distribution functions. In the other limiting case, when tunneling is stronger than scattering (\( \Omega \gg \nu_e \)), the electron is 'collectivized' by the two layers, and the density matrix \( \rho^{(0)} \) is diagonal in the basis of \( H_0 \)-eigenstates. For the parameters used in our calculations, \( h\Omega \approx 10 \) meV exceeds the relaxation rate \( h\nu_e \approx 1 \) meV, and the latter limiting case is justified. Setting \( \rho^{(0)}_{\alpha\beta} = f_\alpha \delta_{\alpha\beta} \), where \( f \)
is the Fermi distribution function, we find
\[
\langle p, s, l | \hat{\rho}^{(1)} | p's'l' \rangle = s_l' \langle n^{p'}_{\omega} \rangle \frac{f^{p's'l'}_{\omega} - f^{p'sl}_{\omega}}{\omega + i\delta} - \langle n^{p'sl}_{\omega} \rangle \frac{f^{p'sl}_{\omega}}{\omega + i\delta},
\]
(A31)