Quantum nonlocal theory of topological Fermi arc plasmons in Weyl semimetals

Gian Marcello Andolina,[1,2,*] Francesco M. D. Pellegrino,[1] Frank H. L. Koppens,[3,4] and Marco Polini[2]

[1]*NEST, Scuola Normale Superiore, I-56126 Pisa, Italy*

[2]*Graphene Labs, Istituto Italiano di Tecnologia, Via Morego 30, I-16163 Genova, Italy*

[3]*ICREA-Institució Catalana de Recerça i Estudis Avancats, 08010 Barcelona, Spain*

[4]*ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels (Barcelona), Spain*

(Received 3 December 2017; published 28 March 2018; corrected 21 June 2019)

The surface of a Weyl semimetal (WSM) displays Fermi arcs, i.e., disjoint segments of a two-dimensional Fermi contour. We present a quantum-mechanical nonlocal theory of chiral Fermi arc plasmons in WSMs with broken time-reversal symmetry. These are collective excitations constructed from topological Fermi arc and bulk electron states and arising from electron-electron interactions, which are treated in the realm of the random phase approximation. Our theory includes quantum effects associated with the penetration of the Fermi arc surface states into the bulk and dissipation, which is intrinsically nonlocal in nature and arises from decay processes mainly involving bulk electron-hole pair excitations.

DOI: 10.1103/PhysRevB.97.125431

I. INTRODUCTION

Plasmons are self-sustained oscillations of the charge density that occur in metals and doped semiconductors [1–3]. The coupling of these matter excitations to photons enables one to squeeze electromagnetic radiation from the visible [4] to the terahertz (THz) [5] spectral range into nanoscale devices. However, plasmons suffer scattering from ever-present extrinsic mechanisms in real solid-state devices [6], including phonons and disorder. Therefore, when confinement is significant, i.e., when the plasmon wavelength λ_p in the material is much smaller than the illumination wavelength $\lambda_0 = 2\pi c/\omega$, losses tend to be high (at room temperature) and hamper potential technological breakthroughs.

Substantial efforts have been recently made to increase the lifetime of these propagating modes at room temperature, without decreasing the associated compression ratio λ_p/λ_0. For example, one can utilize high-quality graphene sheets encapsulated in hexagonal boron nitride [7], where graphene plasmons scatter essentially only against the acoustic phonons of the two-dimensional (2D) carbon lattice [8], which are weakly coupled to the electronic degrees of freedom. Another possible pathway is to use plasmons in topologically nontrivial materials [9–12]. In the particular case of crystals displaying broken time-reversal symmetry (BTRS), the existence of unidirectional propagating modes akin to the ultra-long-lived [13] topological [14] edge magnetoplasmons that occur in 2D electron systems in the quantum Hall regime [3] is expected. Technologically, it would be extremely useful to use materials where BTRS occurs *without* the aid of an external magnetic field. Natural candidates among topological materials with BTRS are recently discovered Weyl semimetals (WSMs) [15–20]. These are semimetals with protected linear band crossings in the Brillouin zone, which act as power-law-decaying sources of Berry curvature [15–20]. Some of these compounds do display intrinsic BTRS [21] and, at the same time, have intriguing topological surface states called "Fermi arcs" (FAs) [15–20]. These are disjoint segments of a 2D Fermi contour—see Fig. 1(b)—which have so far been imaged only with angle-resolved photoemission spectroscopy (ARPES) [22,23]. Provided that losses are not too strong and that the plasmon electric field leaks sufficiently outside the material, the plasmonic excitations of FA surface states can in principle be explored with spatial resolution using scanning-type near-field optical spectroscopy (for a recent review, see, e.g., Ref. [24]).

Theoretically, a few aspects of plasmons in WSMs have been studied for both cases of bulk [25] and surface [26] modes. In particular, the authors of Ref. [25] carried out a study of bulk plasmons in WSMs, by coupling the Maxwell equations as modified by the axion term [27] with a *local* approximation for the constitutive equation, $J_\alpha(\boldsymbol{q},\omega) \approx \sum_\beta \sigma_{\alpha\beta}(\boldsymbol{0},\omega)E_\beta(\boldsymbol{q},\omega)$. The local conductivity tensor $\sigma_{\alpha\beta}(\boldsymbol{0},\omega)$ was calculated [25] by using semiclassical Boltzmann transport theory in the clean limit. The authors of Ref. [26] studied WSM surface plasmon polaritons by using the same approach and deep in the retarded limit, i.e., for surface plasmon wave numbers $|\boldsymbol{q}_\parallel| \sim \omega/c$. Finally, the authors of Ref. [28] carried out a semiclassical study of WSM surface plasmons, which is intrinsically valid in the long-wavelength limit. Nonlocal corrections were heuristically added to the semiclassical equations of motion. An important open question on the plasmon lifetime remains. In order to quantify this, the quantum-mechanical coupling between Fermi arc surface states and bulk states needs to be evaluated.

In this paper, we present a fully quantum-mechanical theory of WSM Fermi arc (FA) plasmons that goes beyond the state of the art [25,26,28]. The present derivation focuses on the simplest microscopic model Hamiltonian of a (type-I) WSM with BTRS [15–19] and is based on linear response theory [1–3] and the random phase approximation (RPA) [1–3]. We

*gian.andolina@sns.it

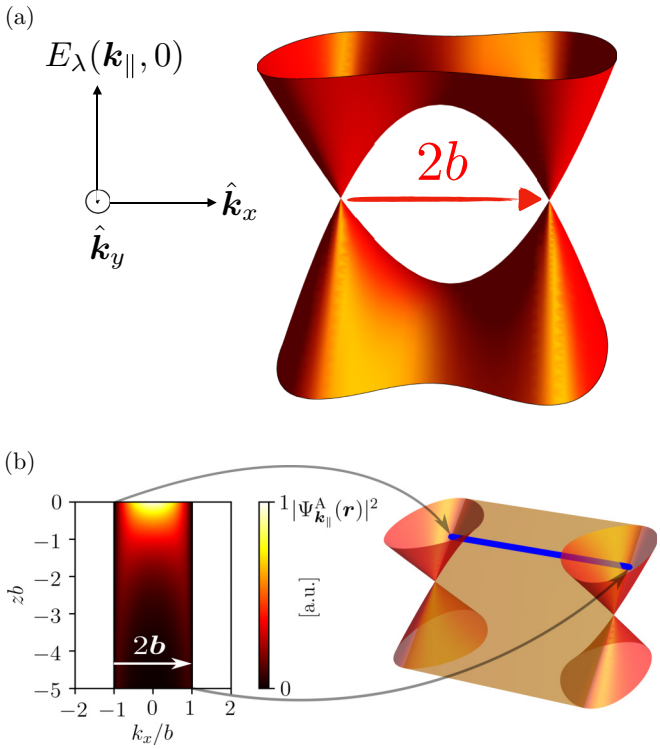

FIG. 1. (a) shows the bulk band structure $E_\lambda(\boldsymbol{k})$ of a WSM with BTRS, for $k_z = 0$. Note that the two Weyl nodes are separated by a vector $2\boldsymbol{b}$ in the $\hat{\boldsymbol{k}}_x \| \hat{\boldsymbol{x}}$ direction. (b) shows the square modulus of the FA wave function in Eq. (3) as a function of the penetration z (in units of $1/b$) into the bulk of a WSM, for a given value of k_x/b. The right-hand side of the panel shows a low-energy zoom of the bulk band structure with two Weyl cones, with the sheet representing the FA dispersion relation $E_{\rm A}(\boldsymbol{k}_\parallel)$.

focus on the electrostatic regime, where the plasmon wave number $|\boldsymbol{q}_\parallel|$ is much larger than the photon one (ω/c), enabling a great concentration of electromagnetic energy. We claim that quantum *nonlocal* effects are crucial to understand WSM FA plasmon physics. First of all, our theory captures analytically the first nonlocal correction to the FA plasmon dispersion in the limit $|\boldsymbol{q}_\parallel| \ll k_{\rm F}$, where $k_{\rm F}$ is the bulk Fermi wave number. Second, since the FA wave functions are in contact (i.e., strong spatial overlap) with a bulk of gapless excitations, FA plasmons are susceptible to Landau damping even at zero temperature and deep in the long-wavelength $|\boldsymbol{q}_\parallel| \ll k_{\rm F}$ limit. We quantify this intrinsic dissipation mechanism, which is dominated by processes whereby FA plasmons decay by emitting electron-hole pairs in the bulk. Our calculations on the FA plasmon lifetime pose strict bounds on the observability of certain angular portions of the highly anisotropic FA plasmon dispersion. Indeed, since FA plasmon modes can be strongly overdamped along certain angular directions, it is essential to consider the dispersion relation as well as the angular dependence of the losses. Finally, we study the pattern of FA plasmon waves that can be measured by carrying out a scattering-type near-field optical experiment (s-SNOM) [24] on the surface of a WSM with BTRS. In this class of experiments, light is focused on a metallized tip with the aim of launching propagating surface plasmons. We calculate the screened potential associated with FA plasmons, which clearly shows characteristic features due to their peculiar propagation dynamics.

Our paper is organized as following. In Sec. II we present the single-particle Hamiltonian and corresponding bulk and surface eigenvalues and eigenstates of a type-I semi-infinite WSM with BTRS. In Sec. III we lay down a general theory of surface plasmons hosted by electron systems occupying a semi-infinite space. In Sec. IV we present a formal expression of the linear density-density response function of a WSM based on the Lehmann representation. We also introduce analytical asymptotic formulas for the same quantity in the relevant limits with respect to the 2D conserved wave vector $\boldsymbol{q}_\parallel$ and frequency ω. Section V is devoted to the FA plasmon dispersion, while intrinsic damping of these modes is discussed in Sec. VI. Finally, in Sec. VII we present a calculation of the spatial pattern of FA plasmon waves that would be seen in an s-SNOM experiment on the surface of a WSM with BTRS. A brief summary is reported in Sec. VIII. Five Appendixes report a number of very useful technical details.

II. SINGLE-PARTICLE PHYSICS OF SEMI-INFINITE WSMs

In order to describe a three-dimensional (3D) WSM with BTRS and two Weyl nodes separated by a vector $2\boldsymbol{b}$ (see Fig. 1), we use the following family of Hamiltonians [29] depending on the parameter m,

$$\mathcal{H}_m(\boldsymbol{k}) = \frac{\hbar v}{2b}\left(k_x^2 - m\right)\sigma_x + \hbar v(k_y\sigma_z + k_z\sigma_y). \tag{1}$$

In Eq. (1), v is the Weyl fermion's velocity and σ_i are ordinary 2×2 Pauli matrices. The first term on the right-hand side of Eq. (1) breaks time-reversal symmetry. For $m = b^2 > 0$, the low-energy spectrum of $\mathcal{H}_m(\boldsymbol{k})$ contains two Weyl nodes separated by $2\boldsymbol{b} = 2b\hat{\boldsymbol{k}}_x$ with $b > 0$. The spectrum of $\mathcal{H}_m(\boldsymbol{k})$ is $E_\lambda(\boldsymbol{k}) = \lambda\hbar v\sqrt{K_x^2 + k_y^2 + k_z^2}$, where $\lambda = \pm 1$ denotes conduction/valance band states and $K_x \equiv (k_x^2 - b^2)/(2b)$. For $k_x \approx \pm b$, $E_\lambda(\boldsymbol{k})$ is linear and isotropic, with two Weyl cones located at $\boldsymbol{k} = \pm b\hat{\boldsymbol{k}}_x$. The corresponding eigenstates are $\Psi_{\lambda,\boldsymbol{k}}(\boldsymbol{r}) = V^{-1/2}u_{\boldsymbol{k},\lambda}e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$, where $V = SL_z$ is the 3D electron system volume, S the surface area, L_z its thickness, and $u_{\boldsymbol{k},\lambda}$ is a spinor given by

$$u_{\boldsymbol{k},\lambda} = \frac{1}{\sqrt{2}}\begin{pmatrix}\sqrt{1+\lambda\cos(\beta_{\boldsymbol{k}})}e^{-i\phi_{\boldsymbol{k}}} \\ \lambda\sqrt{1-\lambda\cos(\beta_{\boldsymbol{k}})}\end{pmatrix}, \tag{2}$$

with $\cos(\beta_{\boldsymbol{k}}) = k_y/\sqrt{K_x^2 + k_y^2 + k_z^2}$, $\sin(\beta_{\boldsymbol{k}}) = \sqrt{(K_x^2 + k_z^2)/(K_x^2 + k_y^2 + k_z^2)}$, and $e^{i\phi_{\boldsymbol{k}}} = (K_x + ik_z)/\sqrt{K_x^2 + k_z^2}$.

In this paper, we are interested in the surface states of a 3D WSM, which we obtain from the following procedure. We first note that, for $m < 0$, $\mathcal{H}_m(\boldsymbol{k})$ describes an insulator with a gap $E_{\rm g} = -\hbar v m/(2b)$. The vacuum can be therefore modeled by taking the $\lim_{m\to-\infty}\mathcal{H}_m(\boldsymbol{k})$. Fermi arcs (FAs) are present only on a surface parallel to [18,30] $2\boldsymbol{b} = 2b\hat{\boldsymbol{k}}_x$. Orienting the crystal in such a way that $(\hat{\boldsymbol{k}}_x, \hat{\boldsymbol{k}}_y, \hat{\boldsymbol{k}}_z) \| (\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}})$, we locate the vacuum/WSM interface hosting the FAs at $z = 0$. Surface states emerge by assuming that the parameter m in Eq. (1) changes sign with z, i.e., $m(z) = b^2\Theta(-z) - \tilde{m}\Theta(z)$,

where $\Theta(z)$ is the Heaviside step function and $\tilde{m}$ is a positive constant. This choice describes a WSM for $z<0$ and the vacuum for $z>0$, provided that one takes the limit $\tilde{m}\to+\infty$. Breaking translational invariance along the $\hat{\boldsymbol{z}}$ direction requires the change $k_z\to -i\partial_z$. FAs are described by states that are bound to the surface. Following Ref. [29], we find that their wave functions are

$$\Psi^{\rm A}_{\boldsymbol{k}_\parallel}(\boldsymbol{r})=\sqrt{\frac{b^2-k_x^2}{bS}}\begin{pmatrix}1\\0\end{pmatrix}\Theta(-z)e^{z/\ell}e^{i\boldsymbol{k}_\parallel\boldsymbol{r}_\parallel}, \tag{3}$$

where $\boldsymbol{k}_\parallel=(k_x,k_y)$, $\ell=2b/(b^2-k_x^2)$, and the index "A" in Eq. (3) stands for "arc." These states have chiral dispersion, $E_{\rm A}(\boldsymbol{k}_\parallel)=\hbar v k_y$, group velocity given by v, density of states given by $\mathcal{N}_{\rm A}=b/(2\pi^2\hbar v)$ (which is independent of the bulk doping), and exist only between the two Weyl nodes, i.e., for $-b\leqslant k_x\leqslant b$. Note that ℓ, which physically represents the extension of the FA state into the WSM bulk $z<0$, goes to infinity for $k_x\to\pm b$. This means that for k_x near the location of the Weyl nodes, leakage of the FA states into the bulk cannot be neglected. For the sake of simplicity, the FA states (3) will be denoted by the shorthand $|{\rm A}\rangle$.

The presence of the WSM/vacuum interface at $z=0$ affects also the propagating (i.e., bulk) states. Since $\boldsymbol{k}_\parallel$ is a good quantum number (because of translational invariance in the $\hat{\boldsymbol{x}}$-$\hat{\boldsymbol{y}}$ plane), we obtain propagating states in the presence of the interface by taking linear combinations of bulk plane-wave states with positive ($k_z>0$) and negative ($-k_z<0$) values of the $\hat{\boldsymbol{z}}$ component of the wave vector $\boldsymbol{k}$. Assuming specular reflection at the interface, we find

$$\Psi^{\rm B}_{\lambda,\boldsymbol{k}}(\boldsymbol{r}_\parallel,z)=\frac{\Theta(-z)}{\sqrt{V}}\left(u_{\boldsymbol{k},\lambda}e^{ik_z z}+r_{\boldsymbol{k}}u_{\bar{\boldsymbol{k}},\lambda}e^{-ik_z z}\right)e^{i\boldsymbol{k}_\parallel\cdot\boldsymbol{r}_\parallel}, \tag{4}$$

where $\bar{\boldsymbol{k}}=(\boldsymbol{k}_\parallel,-k_z)$ is the reflected wave vector and $r_{\boldsymbol{k}}$ is a reflection coefficient (with $|r_{\boldsymbol{k}}|=1$) to be determined by imposing suitable boundary conditions. Of course, $\Psi^{\rm B}_{\lambda,\boldsymbol{k}}(\boldsymbol{r}_\parallel,z)$ is an eigenstate of the half-space Hamiltonian $\lim_{\tilde{m}\to+\infty}\mathcal{H}_{m(z)}(\boldsymbol{k}_\parallel,k_z\to -i\partial_z)$ with eigenvalue $E_\lambda(\boldsymbol{k})$. As explained in Appendix A, the continuity of the wave functions at the interface requires $r_{\boldsymbol{k}}=-1$. Because the "reflected" states (4) originate from the WSM bulk states $\Psi_{\lambda,\boldsymbol{k}}(\boldsymbol{r})$, we will denote them by the shorthand $|{\rm B}\rangle$.

III. SURFACE PLASMONS OF SEMI-INFINITE WSMs

Due to the presence of the interface at $z=0$, it is natural to lay down linear response theory [3] by imposing translational invariance in the direction perpendicular to $\hat{\boldsymbol{z}}$ and that the wave vector $\boldsymbol{q}_\parallel$ parallel to the surface is a good quantum number. The linear density response n_1 induced by an external potential is therefore a function of z, $\boldsymbol{q}_\parallel$, and frequency ω. It can be expressed in terms of the screened potential [3] $V_{\rm sc}(z,\boldsymbol{q}_\parallel,\omega)$ and the proper density-density response function [3] $\tilde{\chi}_{nn}(z,z',\boldsymbol{q}_\parallel,\omega)$ according to

$$n_1(z,\boldsymbol{q}_\parallel,\omega)=\int_{-\infty}^{0}dz'\,\tilde{\chi}_{nn}(z,z',\boldsymbol{q}_\parallel,\omega)V_{\rm sc}(z',\boldsymbol{q}_\parallel,\omega). \tag{5}$$

For a self-sustained oscillation occurring in the absence of an external potential, the screened potential is related to the induced density by

$$V_{\rm sc}(z,\boldsymbol{q}_\parallel,\omega)=\int dz'\int dz''v(z,z',\boldsymbol{q}_\parallel)\tilde{\chi}_{nn}(z',z'',\boldsymbol{q}_\parallel,\omega)\times V_{\rm sc}(z'',\boldsymbol{q}_\parallel,\omega), \tag{6}$$

where the integrals over z' and z'' span the half space $z<0$ and $v(z,z',\boldsymbol{q}_\parallel)=2\pi e^2\exp(-q_\parallel|z-z'|)/q_\parallel$ is the Fourier transform of the Coulomb potential ($q_\parallel=|\boldsymbol{q}_\parallel|$).

Plasmons are nontrivial solutions of the integral equation (6). In this paper, we are solely interested in plasmons whose associated electric field is bound to the WSM/vacuum interface. We therefore set $V_{\rm sc}(z,\boldsymbol{q}_\parallel,\omega)=\bar{v}_{\rm sc}(\boldsymbol{q}_\parallel,\omega)e^{-q_\parallel|z|}$, which describes a mode bound to the surface with a "localization" length scale equal to $1/q_\parallel$. Utilizing this ansatz for $V_{\rm sc}(z,\boldsymbol{q}_\parallel,\omega)$, we evaluate [31] Eq. (6) for $z>0$. We find that this integral equation can be written as a standard algebraic 2D plasmon equation,

$$1=\frac{2\pi e^2}{q_\parallel}\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega), \tag{7}$$

provided that one introduces the effective 2D proper response function,

$$\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)\equiv\int_{-\infty}^{0}dz\int_{-\infty}^{0}dz'\,\tilde{\chi}_{nn}(z,z',\boldsymbol{q}_\parallel,\omega)e^{q_\parallel(z+z')}=L_z\tilde{\chi}_{nn}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)\big|_{q_z=iq_\parallel,q_z'=-iq_\parallel}. \tag{8}$$

Here,

$$\tilde{\chi}_{nn}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)\equiv\frac{1}{L_z}\int\frac{dq_z}{2\pi}\int\frac{dq_z'}{2\pi}\tilde{\chi}_{nn}(z,z',\boldsymbol{q}_\parallel,\omega)\times e^{-iq_z z}e^{+iq_z' z'}. \tag{9}$$

No approximation has yet been made on $\tilde{\chi}_{nn}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)$. The only key assumption we used to derive Eq. (7) was that $\tilde{\chi}_{nn}(z,z',\boldsymbol{q}_\parallel,\omega)\propto\Theta(-z)\Theta(-z')$.

Screening due to the dielectric background can be taken care of by replacing $e^2\to e^2/\bar{\epsilon}$ in Eq. (7), where $\bar{\epsilon}\equiv(1+\epsilon_{\rm b})/2$ and $\epsilon_{\rm b}$ is the high-frequency WSM bulk dielectric constant. This prescription, which is valid only for surface modes, is demonstrated in Appendix B.

IV. NONINTERACTING RESPONSE FUNCTION OF A SEMI-INFINITE WSM

In the RPA [3], the exact proper response function $\tilde{\chi}_{nn}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)$ in Eq. (9) is replaced by the noninteracting response function $\chi^{(0)}_{nn}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)$, which can be evaluated with the aid of the Lehmann representation [3],

$$\chi^{(0)}_{nn}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)=\frac{1}{V\hbar}\sum_{m,n}\frac{f_n-f_m}{\omega_{n,m}+\omega+i0^+}\langle n|\hat{n}_{\boldsymbol{q}_\parallel,q_z}|m\rangle\times\langle m|\hat{n}_{-\boldsymbol{q}_\parallel,-q_z'}|n\rangle\,. \tag{10}$$

Here, $|n\rangle$, $|m\rangle$ (E_n,E_m) denote the eigenstates (eigenenergies) of the noninteracting Hamiltonian, $\hbar\omega_{n,m}=E_n-E_m$ are the excitation energies, f_n are the occupation numbers, and $\hat{n}_{\boldsymbol{q}_\parallel,q_z}\equiv e^{-i\boldsymbol{q}_\parallel\cdot\hat{\boldsymbol{r}}_\parallel-iq_z\hat{z}}$. As we have seen above, the states of our semi-infinite WSM comprise the FA states (3) and the

reflected states (4). We can therefore naturally decompose $\chi^{(0)}_{nn}(q_z,q'_z,\boldsymbol{q}_\parallel,\omega)$ into the sum of three contributions,

$$\chi^{(0)}_{nn} = \chi^{(0)}_{\rm AA} + \chi^{(0)}_{\rm BB} + \chi^{(0)}_{\rm AB}. \tag{11}$$

In the first contribution, $\chi^{(0)}_{\rm AA}$, the sum over n and m in Eq. (10) spans only the FA states (3), i.e., $|n\rangle = |{\rm A}\rangle$, $|m\rangle = |{\rm A}\rangle'$. In the second contribution, $\chi^{(0)}_{\rm BB}$, the sum spans only the reflected states (4), i.e., $|n\rangle = |{\rm B}\rangle$, $|m\rangle = |{\rm B}\rangle'$. Finally, the third contribution, $\chi^{(0)}_{\rm AB}$, takes into account the remaining cross processes in which $|n\rangle = |{\rm A}\rangle$ and $|m\rangle = |{\rm B}\rangle$ or $|n\rangle = |{\rm B}\rangle$ and $|m\rangle = |{\rm A}\rangle$. Because of Eq. (8), the same decomposition holds true for $\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)$, i.e., $\chi_{\rm eff} = \chi^{\rm eff}_{\rm AA} + \chi^{\rm eff}_{\rm BB} + \chi^{\rm eff}_{\rm AB}$. The FA plasmon dispersion is controlled by ${\rm Re}[\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)]$, while the plasmon damping rate depends on dissipation, i.e., ${\rm Im}[\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)]$.

To make analytical progress, we concentrate our attention on FA plasmons in the long-wavelength limit, defined by the regime in which $q_\parallel$ is the smallest wave-vector scale in the problem. Equation (8) implies $q_z \sim q'_z \sim q_\parallel$. We denote by the shorthand q the small quantity $q_z \sim q'_z \sim q_\parallel$.

In the limit $q \ll b,\omega/v$, we find that the leading-order FA effective response function reduces to (Appendix C)

$${\rm Re}\big[\chi^{\rm eff}_{\rm AA}\big] \to \frac{q_\parallel\cos(\theta)}{\hbar\omega(2\pi)^2}\left[2b\left(1+\frac{vq_\parallel\cos(\theta)}{\omega}\right) - q_\parallel|\sin(\theta)|\right]. \tag{12}$$

As the FA density of states $\mathcal{N}_{\rm A}$, ${\rm Re}[\chi^{\rm eff}_{\rm AA}]$ does not depend on the bulk carrier density n. Neglecting for a moment the contribution to Eq. (7) from the reflected states, ${\rm Re}[\chi^{\rm eff}_{\rm BB}]$, and the cross term, ${\rm Re}[\chi^{\rm eff}_{\rm AB}]$, we find a long-wavelength FA plasmon dispersion

$$\Omega^{\rm FA}_\theta = \frac{\alpha_{\rm ee} vb}{\pi}\cos(\theta), \tag{13}$$

where θ is the angle between $\boldsymbol{q}_\parallel$ and the $\hat{\boldsymbol{y}}$ axis (i.e., the direction along which the FA states have finite group velocity), $\cos(\theta) = q_y/q_\parallel$, and $\alpha_{\rm ee} = e^2/(\hbar v\bar{\epsilon})$ is a dimensionless coupling constant describing the strength of electron-electron interactions. The strong directionality dependence of $\Omega^{\rm FA}_\theta$ is evident. Interestingly, the group velocity $\boldsymbol{v}^{\rm FA}_\theta = \hat{\theta}q_\parallel^{-1}\partial_\theta\Omega^{\rm FA}_\theta = -\alpha_{\rm ee}vb\sin(\theta)\hat{\theta}/(\pi q_\parallel)$ of the mode in Eq. (13) vanishes when the plasmon wave vector $\boldsymbol{q}_\parallel$ is in the $\hat{\boldsymbol{y}}$ direction.

We now turn to evaluate the contribution $\chi^{(0)}_{\rm BB}(q_z,q'_z,\boldsymbol{q}_\parallel,\omega)$ due to the reflected states (4). For a Fermi wave number $k_{\rm F} \ll b$, we can linearize the spectrum $E_\pm(\boldsymbol{k})$ around the points $k^{(j)}_x = jb$, where $j = \pm 1$ is the Weyl cone index. Again, we work in the long-wavelength limit, which in this case is defined by $q_\parallel \ll \omega/v,k_{\rm F}$. Noting that $q_z,q'_z \sim q_\parallel$, we can state that $q_\parallel$ is the smallest wave vector in the problem. Moreover, we assume to be deep in the single-particle optical absorption gap, i.e., we consider $\hbar\omega \ll 2E_{\rm F}$, where $E_{\rm F} = \hbar v k_{\rm F}$ is the Fermi energy. In this regime, interband contributions to all response functions involving the reflected states (4) can be neglected and we can safely use the high-frequency moment expansion [3,32] for the intraband contribution to ${\rm Re}[\chi^{(0)}_{\rm BB}]$. As detailed in Appendix D, we find

$${\rm Re}\big[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)\big] \to \frac{nq_\parallel}{m_{\rm eff}\omega^2}, \tag{14}$$

where $n = k_{\rm F}^3/(6\pi^2)$ is the density of bulk electrons and $m_{\rm eff} = E_{\rm F}/v^2$ the effective mass. Corrections to Eq. (14) scale as $q_\parallel^3$, while terms $\propto q_\parallel^2$ are exactly zero. Notice that inserting only ${\rm Re}[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)]$ in Eq. (7), we recover the well-known Ritchie relation [33] $\Omega_{\rm s} = \Omega_{\rm b}\sqrt{\epsilon_{\rm b}/(2\bar{\epsilon})} = vk_{\rm F}\sqrt{\alpha_{\rm ee}/(3\pi)}$ for the surface plasmon frequency, $\Omega^2_{\rm b} = 4\pi ne^2/(\epsilon_{\rm b}m_{\rm eff})$ being the square of the bulk plasmon frequency. In the undoped $k_{\rm F} = 0$ limit, Eq. (14) yields ${\rm Re}[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)] = 0$.

Finally, we need to evaluate the cross contributions $\chi^{(0)}_{\rm AB}(q_z,q'_z,\boldsymbol{q}_\parallel,\omega)$ and ${\rm Re}[\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)]$. Up to leading order in the long-wavelength limit and, as above, deep in the single-particle optical absorption gap $\hbar\omega/(2E_{\rm F}) \ll 1$, we find (Appendix E)

$${\rm Re}\big[\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)\big] \to -\frac{q_\parallel^2[1+\sin^2(\theta)]}{3\pi^2}\frac{1}{\sqrt{2E_{\rm F}|\hbar\omega|}}. \tag{15}$$

Clearly, Eq. (15) cannot be used in the undoped $k_{\rm F} \to 0$ limit. In this case an explicit calculation (Appendix E) yields $\lim_{k_{\rm F}\to 0}{\rm Re}[\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)] = 0$. Note that $\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)$ does not depend on b to leading order for $q_\parallel \ll b$. In this limit, indeed, $\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)$ describes electron-hole transitions near each Weyl node.

V. FA PLASMON DISPERSION

We now solve Eq. (7) in the RPA, i.e., by replacing ${\rm Re}[\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)]$ with ${\rm Re}[\chi^{\rm eff}_{\rm AA}(\boldsymbol{q}_\parallel,\omega) + \chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega) + \chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)]$. In the following, we focus only on positive frequencies, since modes with negative frequency can be obtained from the relation $\Omega(\boldsymbol{q}_\parallel) = -\Omega(-\boldsymbol{q}_\parallel)$.

For doped WSMs ($k_{\rm F} \neq 0$) we find that the plasmon dispersion in the long-wavelength limit is given by $\Omega(\boldsymbol{q}_\parallel) = \Omega_\theta + \delta\Omega(\boldsymbol{q}_\parallel)$, where Ω_θ is the value of the plasmon frequency for $q_\parallel \to 0$ and $\delta\Omega(\boldsymbol{q}_\parallel)$ is the first-order nonlocal correction in $q_\parallel$,

$$\Omega_\theta = \frac{1}{2}\left[\Omega^{\rm FA}_\theta + \sqrt{\left(\Omega^{\rm FA}_\theta\right)^2 + 4\Omega_{\rm s}^2}\right] \tag{16}$$

and

$$\delta\Omega(\boldsymbol{q}_\parallel) = \alpha_{\rm ee}vq_\parallel\mathcal{I}(\theta). \tag{17}$$

In Eq. (17) we have introduced the quantity

$$\mathcal{I}(\theta) = \frac{\Omega_\theta}{\sqrt{\left(\Omega^{\rm FA}_\theta\right)^2 + 4\Omega_{\rm s}^2}}\left\{\frac{\cos(\theta)}{2\pi}\left[\frac{2bv\cos(\theta)}{\Omega_\theta} - |\sin(\theta)|\right] - \frac{2[1+\sin^2(\theta)]}{3\pi}\sqrt{\frac{\hbar\Omega_\theta}{2E_{\rm F}}}\right\}. \tag{18}$$

The first line in the previous equation is due to ${\rm Re}[\chi^{\rm eff}_{\rm AA}(\boldsymbol{q}_\parallel,\omega)]$, while the second line is due to ${\rm Re}[\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)]$. Equation (16) is formally identical [26] to the dispersion relation of a surface magnetoplasmon [34] in the case of a magnetic field oriented along $\boldsymbol{b}$ and with $\Omega^{\rm FA}_\theta \to \omega_{\rm c}\cos(\theta)$, $\omega_{\rm c}$ being the cyclotron frequency. This result reflects the fact that the topological FA states play the same role of edge states in the quantum Hall regime. The FA plasmon dispersion relation $\Omega(\boldsymbol{q}_\parallel)$ is illustrated in Figs. 2(a) and 3(a). In Fig. 2(a) we clearly note that, for every value of q_x, there is a large asymmetry in the dispersion $\Omega(\boldsymbol{q}_\parallel)$

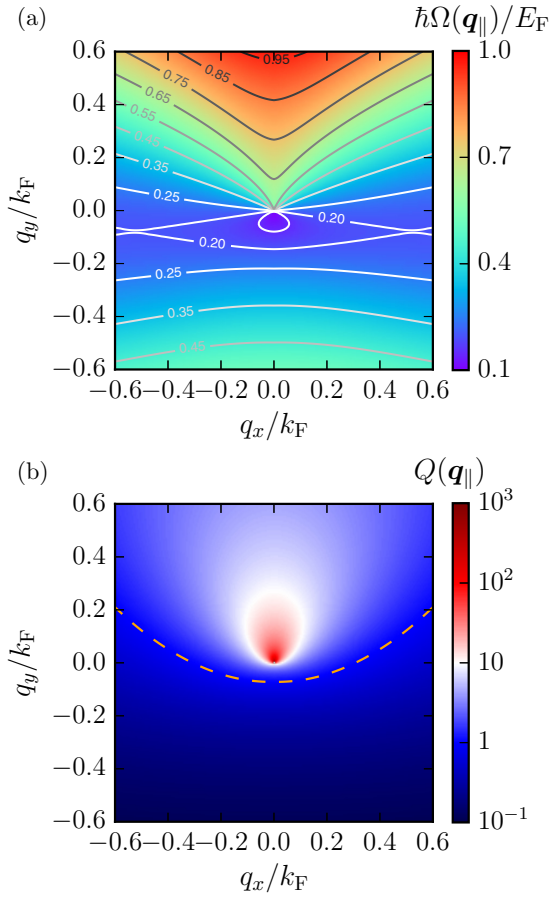

FIG. 2. (a) displays the FA plasmon dispersion relation $\Omega(\boldsymbol{q}_\parallel)$ in units of $E_F/\hbar$ and as a function of q_x/k_F and q_y/k_F. The plasmon group velocity $\boldsymbol{v}_g$ is orthogonal to the contour lines. (b) shows the quality factor $Q(\boldsymbol{q}_\parallel)$ as a function of q_x/k_F and q_y/k_F. The orange dashed line represents $Q(\boldsymbol{q}_\parallel) = 1$. Below this line the plasmon is not a well-defined excitation. In both panels we have set $\alpha_{ee} = 0.5$ and $b = 3k_F$.

as a function of q_y. This stems from the fact that, for $q_y > 0$, $\Omega(\boldsymbol{q}_\parallel)$ is dominated by the FA states. On the other hand, for $q_y < 0$, the main contribution to $\Omega(\boldsymbol{q}_\parallel)$ is given by Ω_s, which stems from the reflected states (4) and is weakly dispersive. The weak dispersive features seen for $q_y < 0$ reflect a hybridization between the FA and bulk channels. This mixing of channels with very different symmetries (the former highly directional, while the latter is isotropic) leads to a pair of saddle points in the dispersion relation, which are seen from the contours in Fig. 2(a) at $\hbar\Omega(\boldsymbol{q}_\parallel)/E_F \approx 0.20$. In Fig. 3(a) we show cuts of $\Omega(\boldsymbol{q}_\parallel)$, each one taken at a fixed value of $q_\parallel$, and plotted as a function of $-\pi \leqslant \theta \leqslant \pi$.

For typical experimental values [20], $\epsilon_b = 10$, $E_F = 40$ meV, $b = 0.05$ Å^{-1}, and $v = c/1000$, Eq. (16) yields a FA plasmon energy Ω_θ in the forward $\theta \approx 0$ direction on the order of 30 meV, corresponding to a frequency $\nu_\theta \equiv \Omega_\theta/(2\pi)$ of ~7.3 THz.

In the undoped $k_F \to 0$ limit the plasmon dispersion is

$$\Omega(\boldsymbol{q}_\parallel) = \Omega_\theta^{FA} + q_\parallel\left[v\cos(\theta) - |\sin(\theta)|\frac{\Omega_\theta^{FA}}{2b}\right]. \tag{19}$$

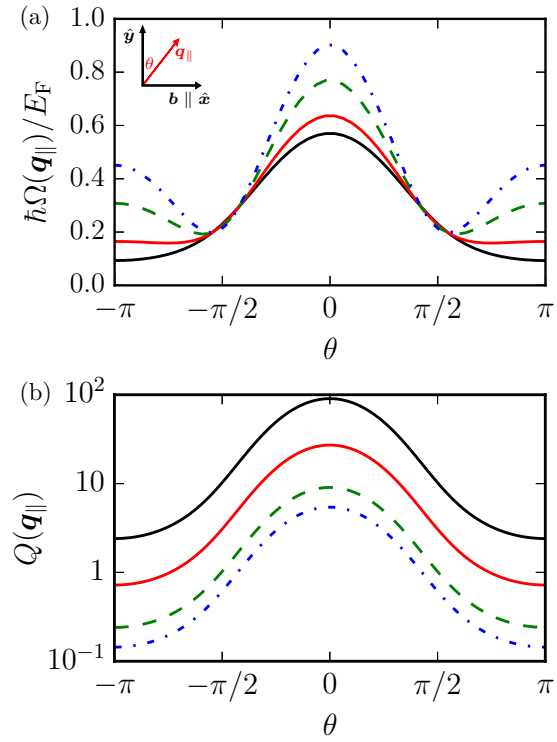

FIG. 3. (a) The plasmon frequency $\Omega(\boldsymbol{q}_\parallel)$ in units of $E_F/\hbar$ as a function of θ for different values of $q_\parallel = |\boldsymbol{q}_\parallel|$ (black solid line: $q_\parallel = 0$; red solid line: $q_\parallel = 0.1k_F$; green dashed line: $q_\parallel = 0.3k_F$; blue dashed-dotted line: $q_\parallel = 0.5k_F$). The inset shows the definition of θ, i.e., the angle between the group velocity of the FA states and the wave vector $\boldsymbol{q}_\parallel$. (b) The dimensionless plasmon quality factor $Q(\boldsymbol{q}_\parallel)$ as a function of θ for different values of $q_\parallel$ (black solid line: $q_\parallel = 0.03k_F$; red solid line: $q_\parallel = 0.1k_F$, green dashed line: $q_\parallel = 0.3k_F$; blue dashed-dotted line: $q_\parallel = 0.5k_F$). In both panels we have set $\alpha_{ee} = 0.5$ and $b = 3k_F$.

We will momentarily show that in the undoped limit the FA plasmon is strongly damped, but for nonzero k_F, this damping is suppressed.

VI. INTRINSIC LIFETIME OF FA PLASMONS

So far, we have treated the plasmon as a solution of Eq. (6) occurring on the real frequency axis. For fundamental reasons related to causality [3], however, retarded response functions must have poles located below the real axis, i.e., at $\omega = \Omega(\boldsymbol{q}_\parallel) - i\Gamma(\boldsymbol{q}_\parallel)$ with $\Gamma(\boldsymbol{q}_\parallel) > 0$. When the imaginary part is small, i.e., when $\Gamma(\boldsymbol{q}_\parallel) \ll \Omega(\boldsymbol{q}_\parallel)$, the plasmon is a well-defined collective excitation of the many-body system [3]. The plasmon lifetime $\tau(\boldsymbol{q}_\parallel)$ is $\Gamma^{-1}(\boldsymbol{q}_\parallel)$. Expanding Eq. (7) in the small parameter $\Gamma(\boldsymbol{q}_\parallel)/\Omega(\boldsymbol{q}_\parallel)$, we obtain

$$\Gamma(\boldsymbol{q}_\parallel) = \frac{\mathrm{Im}[\chi_{\mathrm{eff}}(\boldsymbol{q}_\parallel,\Omega(\boldsymbol{q}_\parallel))]}{\partial_\omega\,\mathrm{Re}[\chi_{\mathrm{eff}}(\boldsymbol{q}_\parallel,\omega)]|_{\omega=\Omega(\boldsymbol{q}_\parallel)}}. \tag{20}$$

We define the quality factor as $Q(\boldsymbol{q}) \equiv \Omega(\boldsymbol{q}_\parallel)/[2\Gamma(\boldsymbol{q}_\parallel)]$. We first calculate $\partial_\omega\,\mathrm{Re}[\chi_{\mathrm{eff}}(\boldsymbol{q}_\parallel,\omega)|_{\omega=\Omega(\boldsymbol{q}_\parallel)}]$ at leading order in $q_\parallel$. Only $\chi^{(0)}_{AA}$ and $\chi^{(0)}_{BB}$ contribute to this quantity at leading order in $q_\parallel$. We find $\partial_\omega\,\mathrm{Re}[\chi_{\mathrm{eff}}(\boldsymbol{q}_\parallel,\omega)|_{\omega=\Omega(\boldsymbol{q}_\parallel)}] \to -q_\parallel[\cos(\theta)b/2 + vk_F^2/(3\Omega_\theta)]/(\hbar\Omega_\theta^2\pi^2)$. We then calculate $\mathrm{Im}[\chi_{\mathrm{eff}}(\boldsymbol{q}_\parallel,\omega)]$. As detailed in Appendix C, $\mathrm{Im}[\chi^{\mathrm{eff}}_{AA}(\boldsymbol{q}_\parallel,\omega)] \to$

$-q_\parallel b\cos(\theta)\delta(\omega - E_{\rm A}(\boldsymbol{q}_\parallel)/\hbar)/(2\hbar\pi)$. This result can be easily understood as the absorption spectrum of a one-dimensional system [3], whose role in our case is effectively played by the FA states. For the calculation of $\mathrm{Im}[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)]$, we use the well-known approximate relation [35] $\chi^{\rm eff}_{\rm BB}(z,z',\boldsymbol{q}_\parallel,\omega) \simeq \Theta(-z)\Theta(-z')[\chi^{\rm (h)}_{\rm BB}(z+z',\boldsymbol{q}_\parallel,\omega) + \chi^{\rm (h)}_{\rm BB}(z-z',\boldsymbol{q}_\parallel,\omega)]$. Here, $\chi^{\rm (h)}_{\rm BB}(z,\boldsymbol{q}_\parallel,\omega) = \int dq_z \chi^{\rm (h)}_{\rm BB}(\boldsymbol{q}_{\rm 3D},\omega)\exp(iq_z z)/(2\pi)$ is the Fourier transform of the homogeneous response function $\chi^{\rm (h)}_{\rm BB}(\boldsymbol{q}_{\rm 3D},\omega)$ of a bulk WSM [36,37] and $\boldsymbol{q}_{\rm 3D} = (q_z,\boldsymbol{q}_\parallel)$. Straightforward algebraic manipulations yield

$$\mathrm{Im}\big[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)\big] \simeq \int \frac{dq_z}{2\pi}\frac{2q_\parallel^2}{\left(q_\parallel^2+q_z^2\right)^2}\mathrm{Im}\big[\chi^{\rm (h)}_{\rm BB}(\boldsymbol{q}_{\rm 3D},\omega)\big]. \quad (21)$$

Equation (21) is crucial as it states that, for every ω, the decay rate (20) of the FA plasmon depends on the spectral density of electron-hole pairs $\mathrm{Im}[\chi^{\rm (h)}_{\rm BB}(\boldsymbol{q}_{\rm 3D},\omega)]$ in the bulk integrated over all values of q_z. Physically, this equation expresses the fact that a surface plasmon can decay without conserving the $\hat{\boldsymbol{z}}$ component of the wave vector $\boldsymbol{q}_{\rm 3D}$, for the presence of the surface breaks translational invariance in this direction relaxing the three-dimensional momentum conservation that a bulk plasmon would obey. The latter is encoded in regions of the plane $\boldsymbol{q}_{\rm 3D}$-ω where $\mathrm{Im}[\chi^{\rm (h)}_{\rm BB}(\boldsymbol{q}_{\rm 3D},\omega)]$ is zero. Because of the convolution in Eq. (21), this ceases to be true for $\mathrm{Im}[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)]$.

In the limit $\omega \gg v q_\parallel$ and at the leading order in $2E_{\rm F}/(\hbar\omega) \gg 1$, we obtain

$$\mathrm{Im}\big[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)\big] \to -\frac{q_\parallel^2}{32\pi^2\hbar\omega}\left(\frac{2E_{\rm F}}{\hbar\omega}\right)^2. \quad (22)$$

In the limit $k_{\rm F}\to 0$, we find $\mathrm{Im}[\chi^{\rm eff}_{\rm BB}(\boldsymbol{q}_\parallel,\omega)] = -q_\parallel/(24\pi\hbar v)$. This implies that, in the undoped limit, the decay rate Γ tends to a finite value in the long-wavelength $q_\parallel \to 0$ limit. The FA plasmon is therefore not a well-defined excitation in the case of an undoped WSM, since it easily decays by emitting interband electron-hole pairs in the gapless bulk.

Finally, we need to calculate $\mathrm{Im}[\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)]$. Following the steps described in Appendix E, we find

$$\mathrm{Im}\big[\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)\big] \to -\frac{q_\parallel^2\left[1+\sin^2(\theta)\right]}{3\pi^2\hbar\omega}\left(\frac{2E_{\rm F}}{\hbar\omega}\right)^{-1/2}. \quad (23)$$

We emphasize that the contribution to the decay rate coming from Eq. (23) is suppressed with respect to that coming from Eq. (22) by a factor $[2E_{\rm F}/(\hbar\omega)]^{-5/2} \ll 1$. We will therefore neglect the contribution (23) for the calculation of Γ. Physically, Eq. (23) represents processes in which a FA plasmon decays by emitting a composite electron-hole pair, with one partner of the pair belonging to the FA manifold of states and the other partner to the bulk manifold. In the limit $k_{\rm F}\to 0$, we get $\mathrm{Im}[\chi^{\rm eff}_{\rm AB}(\boldsymbol{q}_\parallel,\omega)] = -2q_\parallel^2[1+\sin^2(\theta)]/(3\pi^2\hbar\omega)$.

In summary, for a doped WSM, we get the following compact expression for the decay rate of a FA plasmon in the long-wavelength limit and deep in the single-particle optical gap $\hbar\omega \ll 2E_{\rm F}$,

$$\Gamma(\boldsymbol{q}_\parallel) = \frac{\Omega_\theta}{32}\frac{q_\parallel}{k_{\rm F}}\frac{1}{\frac{\cos(\theta)}{2}\frac{b}{k_{\rm F}} + \frac{1}{6}\left(\frac{2E_{\rm F}}{\hbar\Omega_\theta}\right)}\left(\frac{2E_{\rm F}}{\hbar\Omega_\theta}\right)^2. \quad (24)$$

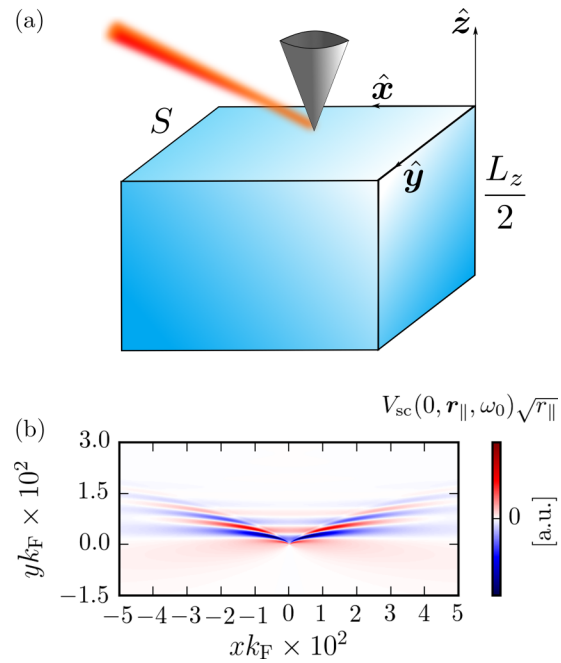

FIG. 4. (a) shows the interface, located on the $\hat{\boldsymbol{x}}$-$\hat{\boldsymbol{y}}$ plane, between a semi-infinite WSM and vacuum. The WSM sample (blue) is a parallelepiped with surface area S and height $L_z/2$. The (gray) cone represents the metallized tip of an atomic force microscope used in an s-SNOM experiment. The latter is illuminated by a laser (red) and located at the origin of the $\hat{\boldsymbol{x}}$-$\hat{\boldsymbol{y}}$ plane. In (b) we present the screened potential in response to a laser with frequency ω_0. The color map shows the calculated screened potential evaluated at $z=0$ and multiplied by the square root of the distance $r_\parallel$ from the tip located at the origin, i.e., $V_{\rm sc}(0,\boldsymbol{r}_\parallel,\omega)\sqrt{r_\parallel}$. Results shown in this panel have been obtained by setting $b/k_{\rm F}=5$, $\alpha_{\rm ee}=0.8$, and $\hbar\omega/E_{\rm F}=1.4$.

The quantity $\Gamma(\boldsymbol{q}_\parallel)/\Omega(\boldsymbol{q}_\parallel)$ is plotted in Figs. 2(b) and 3(b). In writing Eq. (24) we have omitted the contribution $\propto\delta(\omega - E_{\rm A}(\boldsymbol{q}_\parallel)/\hbar)$ stemming from $\mathrm{Im}[\chi^{\rm eff}_{\rm AA}]$. This δ-function contribution is not present in the range of values of $\boldsymbol{q}_\parallel$ we have used to make the plot in Fig. 2(b). From Fig. 2(b) we immediately see that the WSM surface plasmons are highly directional and weakly damped only in the direction of propagation of the single-particle FA states. From Figs. 2(b) and 3(b) we clearly see that damping increases (quality factor decreases) very rapidly as a function of the angle θ.

Before concluding, we would like to comment on one evident limitation of the WSM model in Eq. (1). As discussed above, in this model the FA states disperse only in the $\hat{\boldsymbol{k}}_y$ direction and their dispersion is strictly linear, $E_{\rm A}(\boldsymbol{k}_\parallel) = \hbar v k_y$. More complicated models—see, for example, Ref. [38]—allow of course for more general FA dispersion relations. For example, taking $E_{\rm A}(\boldsymbol{k}_\parallel) = \hbar v k_y + \hbar^2(k_x^2 - b^2)/(2m_x)$, one can demonstrate that Eq. (16) is not modified by the parabolic correction. Also, one can show that the first nonlocal correction in Eq. (17) is modified only for $\theta \sim \pm\pi/2$, i.e., along the directions parallel or antiparallel to the Weyl-node separation vector $2\boldsymbol{b}$. Along these directions, however, the FA plasmon is anyway strongly damped—see Fig. 3(b)—and therefore quantitative changes to its dispersion relation for $\theta \sim \pm\pi/2$ are uninteresting.

VII. RESPONSE OF FA PLASMONS TO AN ILLUMINATED METALLIZED TIP

In this section we present an elementary theory that describes the potential that one would observe by carrying out a scattering-type near-field optical experiment (s-SNOM) [24] on the surface of a WSM with BTRS. In this experiment [24], light is focused on a metallic tip with the aim of launching propagating surface plasmons. In Fig. 4(a) we sketch the setup of such an experiment.

We model the tip as an external potential oscillating in time and with an exponential profile in the $\hat{z}$ direction,

$$v_{\rm ext}(z,\boldsymbol{q}_\parallel,t) = V_0 e^{-q_\parallel |z|} e^{-i\omega_0 t}, \tag{25}$$

where ω_0 is the frequency of the laser that illuminates the tip. The exponential dependence on z dramatically simplifies the algebra. In the Fourier transform, we find $v_{\rm ext}(z,\boldsymbol{q}_\parallel,\omega) \sim \delta(\omega-\omega_0)$. Since we are interested in what happens at the surface of a WSM, we set $z=0$. In this case, the external potential reduces to a delta function of $\boldsymbol{r}_\parallel$, $v_{\rm ext}(0,\boldsymbol{r}_\parallel,t) = V_0\delta(\boldsymbol{r}_\parallel)e^{-i\omega_0 t}$.

Working in the RPA, we can easily write the relationship between $V_{\rm sc}(0,\boldsymbol{q}_\parallel,\omega)$ and the external potential $v_{\rm ext}(0,\boldsymbol{q}_\parallel,\omega)$, which involves the effective 2D response function $\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega_0)$,

$$V_{\rm sc}(0,\boldsymbol{q}_\parallel,\omega_0) = \frac{V_0}{1-\dfrac{2\pi e^2}{q_\parallel}\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega_0)}. \tag{26}$$

Near resonance, i.e., when the laser frequency ω_0 is close to the WSM FA plasmon frequency $\Omega(\boldsymbol{q}_\parallel)$, we can write

$$V_{\rm sc}(0,\boldsymbol{q}_\parallel,\omega_0) \simeq \frac{V_0}{-\dfrac{2\pi e^2}{q_\parallel}\left(\dfrac{\partial\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)}{\partial\omega}\right)_{\omega=\Omega(\boldsymbol{q}_\parallel)}[\omega_0-\Omega(\boldsymbol{q}_\parallel)+i\Gamma(\boldsymbol{q}_\parallel)]}. \tag{27}$$

Fourier transforming the previous result to real space, we find

$$V_{\rm sc}(0,\boldsymbol{r}_\parallel,\omega_0) \simeq V_0\int\frac{d^2\boldsymbol{q}}{(2\pi)^2}\frac{e^{i\boldsymbol{q}_\parallel\boldsymbol{r}_\parallel}}{-\dfrac{2\pi e^2}{q_\parallel}\left(\dfrac{\partial\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)}{\partial\omega}\right)_{\omega=\Omega(\boldsymbol{q}_\parallel)}[\omega_0-\Omega(\boldsymbol{q}_\parallel)+i\Gamma(\boldsymbol{q}_\parallel)]}. \tag{28}$$

This integral can be easily evaluated in polar coordinates,

$$V_{\rm sc}(0,\boldsymbol{r}_\parallel,\omega_{\rm TIP}) \simeq V_0\int_0^\infty q_\parallel dq_\parallel\int_0^{2\pi}d\theta\frac{1}{(2\pi)^2}\frac{e^{iq_\parallel r_\parallel\cos(\theta-\theta')}}{-\dfrac{2\pi e^2}{q_\parallel}\left(\dfrac{\partial\chi_{\rm eff}(q_\parallel,\theta,\omega)}{\partial\omega}\right)_{\omega=\Omega(\boldsymbol{q}_\parallel)}[\omega_0-\Omega(q_\parallel,\theta)+i\Gamma(q_\parallel,\theta)]}, \tag{29}$$

where we have introduced the angle θ' defined by $\boldsymbol{r}_\parallel\cdot\hat{\boldsymbol{y}} = r_\parallel\cos(\theta')$.

Using the result $\partial_\omega{\rm Re}[\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega)|_{\omega=\Omega(\boldsymbol{q}_\parallel)}] \to -q_\parallel[\cos(\theta)b/2+vk_{\rm F}^2/(3\Omega_\theta)]/(\hbar\Omega_\theta^2\pi^2)$ reported above, and noting that the integrand is peaked where the denominator vanishes, we can extend the integration range over $q_\parallel$ from $[0,\infty]$ to $[-\infty,\infty]$. This allows us to use the power of the residue theorem.

The integrand has a simple pole with respect to $q_\parallel$. We define it as $\tilde{p}(\omega_0,\theta) = p_1 + ip_2$, which satisfies the equation $\omega_0 = \Omega(\tilde{p},\theta) - i\Gamma(\tilde{p},\theta)$. Using the definitions of $\mathcal{I}(\theta)$ and Γ_θ given above, we have

$$p_1 = \frac{\omega_0-\Omega_\theta}{\Gamma_\theta^2+[v\alpha\mathcal{I}(\theta)]^2}v\alpha\mathcal{I}(\theta) \tag{30}$$

and

$$p_2 = \frac{\omega_0-\Omega_\theta}{\Gamma_\theta^2+[v\alpha\mathcal{I}(\theta)]^2}\Gamma_\theta. \tag{31}$$

After straightforward mathematical manipulations, we find

$$V_{\rm sc}(0,\boldsymbol{r}_\parallel,\omega_0) \simeq V_0\int_{\theta'-\pi/2}^{\theta'+\pi/2}d\theta\frac{\tilde{p}(\omega_0,\theta)\Omega_\theta^2}{\cos(\theta)b/2+vk_{\rm F}^2/(3\Omega_\theta)} \times\exp[i\,\tilde{p}(\omega_0,\theta)r_\parallel\cos(\theta-\theta')], \tag{32}$$

which needs to be evaluated numerically.

Figure 4(b) displays the calculated screened potential for the following parameters: $b/k_{\rm F}=5$, $\alpha_{\rm ee}=0.8$, and $\hbar\omega/E_{\rm F}=1.4$. Here, it is possible to recognize features due to the peculiar propagation dynamics of FA plasmons. The illustrated SNOM pattern is highly anisotropic because of the highly unidirectional character of both FA dispersion and damping. The anisotropy of the dispersion relation—shown in Fig. 2(a)—manifests through the FA plasmon group velocity, which is maximal around the angular directions $\theta\approx\pm\pi/4$. The anisotropy of the plasmon dissipation—see Fig. 2(b)—shows up in the propagation direction of FA plasmons, which predominantly occurs only along the $y>0$ direction.

VIII. SUMMARY AND CONCLUSIONS

In summary, we have presented a quantum-mechanical nonlocal theory of surface plasmons in semi-infinite Weyl semimetals with broken time-reversal symmetry. We have been able to derive a simple analytical formula—see Eqs. (16)–(18)—for the surface plasmon dispersion relation in the electrostatic limit $|\boldsymbol{q}_\parallel|\gg\omega/c$, which takes into account exquisite quantum effects associated with the penetration of the Fermi arc surface states into the gapless bulk. We have also included nonlocal corrections, which were crucial to investigate in a

quantitative manner the surface plasmon damping rate (24), as determined by decay processes involving the excitation of electron-hole pairs.

Our calculations show that the intrinsic damping of topological Fermi arc plasmons is small at small values of the in-plane wave vector $\boldsymbol{q}_\parallel$, mainly in a specific direction, i.e., the direction along which the single-particle Fermi arc states disperse. Scattering-type near-field optical spectroscopy can therefore be used as an alternative to ARPES to carry out spatially resolved investigations of these intriguing chiral modes occurring in Weyl semimetals with broken time-reversal symmetry, in the absence of an external magnetic field.

ACKNOWLEDGMENTS

It is a great pleasure to thank Iacopo Torre and Fabio Taddei for very useful discussions. M.P. wishes to thank Fondazione Istituto Italiano di Tecnologia for financial support. F.H.L.K. acknowledges financial support from the Government of Catalonia through the SGR grant (2014-SGR-1535), and from the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Programme for Centres of Excellence in R&D (SEV-2015-0522), support by Fundacio Cellex Barcelona, CERCA Programme/Generalitat de Catalunya and the Mineco grants Ramón y Cajal (RYC-2012-12281) and Plan Nacional (FIS2013-47161-P and FIS2014-59639-JIN). Furthermore, the research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 696656 "GrapheneCore1" and the ERC consolidator grant (726001, Toponanop) and Mineco Plan Nacional Grant No. 2D-NANOTOP (FIS2016-81044-P). M.P. is extremely grateful for the financial support granted by ICFO during a visit in August 2016.

APPENDIX A: BOUNDARY CONDITIONS

In order to determine the correct boundary conditions, we solve the problem for $z > 0$, evaluating the evanescent states at $z = 0^+$, and then we take the $\lim_{\tilde{m}\to+\infty} \mathcal{H}_{m(z)}(\boldsymbol{k}_\parallel, k_z \to -i\partial_z)$ in the semi-infinite WSM model Hamiltonian introduced in the main text. We find

$$\lim_{\tilde{m}\to+\infty} \Psi_{E,\boldsymbol{k}_\parallel}(\boldsymbol{r}_\parallel, 0^+) \propto \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{i\boldsymbol{k}_\parallel \boldsymbol{r}_\parallel}, \tag{A1}$$

where $\Psi_{E,\boldsymbol{k}_\parallel}(\boldsymbol{r}_\parallel, z)$ is the evanescent solution of the problem for $z > 0$. Imposing the continuity of the wave function at $z = 0$ between the reflected state in Eq. (4) of the main text and the evanescent state (A1), we find $r_{\boldsymbol{k}} = -1$. Note that the square modulus of the reflected state in Eq. (4) is correctly normalized only in the limit $V \to \infty$.

APPENDIX B: DIELECTRIC BACKGROUND SCREENING

We here describe how to include the effect of a dielectric background in Eq. (7) of the main text. We assume that dielectric screening stems from the response of the empty (occupied) states in the bands above (below) the conduction (valence) band, far away in energy from the Weyl crossing.

The total response $\chi_{\rm tot}$ of the system is the sum of two terms: (i) the response of the states described by the Weyl Hamiltonian in Eq. (1) of the main text, denoted by the symbol χ, and (ii) the contribution due to the high-energy bands, denoted by the symbol $\chi_{\rm b}$, i.e.,

$$\chi_{\rm tot} \equiv \chi + \chi_{\rm b}. \tag{B1}$$

For simplicity, we start from the homogeneous (h) case and we focus on a single pair of bands, with one energetically higher than the conduction band and the other one energetically lower than the valence band. By using the Lehmann representation we can write

$$\chi_{\rm b}^{\rm (h)}(\boldsymbol{q}_{\rm 3D}, \omega) = \frac{1}{V} \sum_{\rm L,H} \frac{f_{\rm L} - f_{\rm H}}{\epsilon_{\rm L} - \epsilon_{\rm H} + \hbar\omega} |\mathcal{M}_{\rm L,H}(\boldsymbol{q}_{\rm 3D})|^2 + [\omega \to -\omega, \boldsymbol{q}_{\rm 3D} \to -\boldsymbol{q}_{\rm 3D}], \tag{B2}$$

where the collective labels "L" and "H" refer to states of a given electronic band with lower (higher) energy than the valence (conduction) band, respectively, and $\mathcal{M}_{\rm L,H}(\boldsymbol{q}_{\rm 3D})$ represents a suitable matrix element. Because the band L (H) is deep below (well above) the Fermi level, we replace $f_{\rm L} \to 1$ ($f_{\rm H} \to 0$). Furthermore, we define $\Delta \equiv \min(\epsilon_{\rm H} - \epsilon_{\rm L}) > 0$. Since we are interested in the frequency range $\omega \ll \Delta/\hbar$, we can approximate Eq. (B2) as

$$\chi_{\rm b}^{\rm (h)}(\boldsymbol{q}_{\rm 3D}, \omega) \simeq -\frac{1}{\Delta V} \sum_{\rm L,H} [|\mathcal{M}_{\rm L,H}(\boldsymbol{q}_{\rm 3D})|^2 + |\mathcal{M}_{\rm L,H}(-\boldsymbol{q}_{\rm 3D})|^2]. \tag{B3}$$

In the long-wavelength limit, because of the orthogonality of the states belonging to different bands, we have $|\mathcal{M}_{\rm L,H}(\boldsymbol{q}_{\rm 3D})|^2 \propto \boldsymbol{q}_{\rm 3D}^2$. We can therefore write

$$\chi_{\rm b}^{\rm (h)}(\boldsymbol{q}_{\rm 3D}, \omega) \simeq -\chi_0 \boldsymbol{q}_{\rm 3D}^2, \tag{B4}$$

where χ_0 is a positive constant, independent of $\boldsymbol{q}$ and ω. By exploiting the generality of the above argument, we can express the global contribution to the response function from all electronic bands far away in energy from the Weyl crossing as $\chi_{\rm b}^{(\rm h)}(\boldsymbol{q}_{\rm 3D},\omega) \simeq -\chi_{\rm b}\boldsymbol{q}_{\rm 3D}^2$, where $\chi_{\rm b} > 0$.

In order to recover the plasmon dispersion relation in the homogeneous case, we have to solve the well-known RPA equation [3]

$$\frac{4\pi e^2}{\boldsymbol{q}_{\rm 3D}^2}\chi_{\rm tot}^{(\rm h)}(\boldsymbol{q}_{\rm 3D},\omega) = 1, \tag{B5}$$

or, equivalently,

$$\frac{4\pi e^2}{\epsilon_{\rm b}\boldsymbol{q}_{\rm 3D}^2}\chi^{(\rm h)}(\boldsymbol{q}_{\rm 3D},\omega) = 1, \tag{B6}$$

where $\epsilon_{\rm b} = 1 + 4\pi e^2\chi_{\rm b}$.

We now follow a similar path for the case of surface plasmon modes. We first replace the expression $\chi_{\rm b}(z,z',\boldsymbol{q}_\parallel,\omega) \simeq \Theta(-z)\Theta(-z')[\chi_{\rm b}^{(\rm h)}(z+z',\boldsymbol{q}_\parallel) + \chi_{\rm b}^{(\rm h)}(z-z',\boldsymbol{q}_\parallel,\omega)]$ in Eq. (B1). We then use the result derived earlier, i.e., $\chi_{\rm b}^{(\rm h)}(\boldsymbol{q}_{\rm 3D},\omega) \simeq -\chi_{\rm b}\boldsymbol{q}_{\rm 3D}^2$. Carrying out straightforward algebraic manipulations, we find Eq. (7) in the main text, i.e.,

$$1 = \frac{2\pi e^2}{\bar{\epsilon}q_\parallel}\chi_{\rm eff}(\boldsymbol{q}_\parallel,\omega), \tag{B7}$$

where $\bar{\epsilon} = (1+\epsilon_{\rm b})/2$.

APPENDIX C: THE AA RESPONSE FUNCTION

We here provide more details on the calculation of the response function $\chi_{\rm AA}^{(0)}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)$. Using the FA eigenstates in Eq. (3) of the main text and the Lehmann representation [3]—Eq. (10) of the main text—we find the following exact expression,

$$\chi_{\rm AA}^{(0)}(q_z,q_z',q_x,q_y,\omega) = \frac{1}{(2\pi)^2 L_z}\frac{q_y}{\hbar(\omega+i0^+) - vq_y}\times\int_{-b}^{b-|q_x|} dk_x\mathcal{L}(q_z,q_z',q_x), \tag{C1}$$

where

$$\mathcal{L}(q_z,q_z',q_x) \equiv \left\{\frac{2[b^2-(k_x+|q_x|)^2]}{(b^2-k_x^2)+[b^2-(k_x+|q_x|)^2]-i2bq_z}\right\}\times\left\{\frac{2(b^2-k_x^2)}{(b^2-k_x^2)+[b^2-(k_x+|q_x|)^2]+i2bq_z'}\right\}. \tag{C2}$$

Because in our model the FA states do not disperse along the $\hat{\boldsymbol{x}}$ direction, we have $\chi_{\rm AA}^{(0)}(q_z,q_z',q_x,q_y=0,\omega)=0$. This means that, in our model, an external field can perturb the FA states only if it carries a finite momentum along the $\hat{\boldsymbol{y}}$ direction. Expanding $\mathcal{L}(q_z,q_z',q_x)$ in a power series of q and carrying out the integral in Eq. (C1), we obtain

$$\mathrm{Re}\big[\chi_{\rm AA}^{(0)}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)\big] \to \frac{1}{(2\pi)^2 L_z}\frac{q_y}{\hbar\omega}\left[2b\left(1+\frac{vqy}{\omega}\right)-|q_x|\right]. \tag{C3}$$

From Eq. (C3) one easily obtains $\mathrm{Re}[\chi_{\rm AA}^{\rm eff}(\boldsymbol{q}_\parallel,\omega)]$, as in Eq. (12) of the main text. Similarly, we find

$$\mathrm{Im}\big[\chi_{\rm AA}^{(0)}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)\big] \to -\frac{q_y}{4\pi L_z}(2b-|q_x|)\delta(\omega-vq_y), \tag{C4}$$

from which $\mathrm{Im}[\chi_{\rm AA}^{\rm eff}(\boldsymbol{q}_\parallel,\omega)]$ follows.

APPENDIX D: THE BB RESPONSE FUNCTION

We here report some useful technical details on the calculation of the response function $\chi_{\rm BB}^{(0)}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)$ for the doped case. Since we are interested in studying the long-wavelength limit, we can use the (high-frequency) moment expansion [3,32], i.e.,

$$\mathrm{Re}\big[\chi_{\rm BB}^{(0)}\big](q_z,q_z',\boldsymbol{q}_\parallel,\omega) \to \frac{\langle\hat{n}_{q_z-q_z'}\rangle_{\rm GS}\left(q_\parallel^2+q_zq_z'\right)}{m_{\rm eff}\omega^2 V}, \tag{D1}$$

where $m_{\rm eff} = E_{\rm F}/v^2$ is the effective mass, $\hat{n}_{q_z-q_z'} = \sum_j e^{i(q_z-q_z')\hat{z}_j}$ is the Fourier transform of the density operator [3], and $\langle\cdots\rangle_{\rm GS}$ is a shorthand for the following ground-state expectation value $\langle{\rm GS}|_{\rm B}\cdots|{\rm GS}\rangle_{\rm B}$. Here, $|{\rm GS}\rangle_{\rm B} = \prod_{|\boldsymbol{k}|<k_{\rm F}} b_{+,\boldsymbol{k}}^\dagger|0\rangle$, where $b_{+1,\boldsymbol{k}}^\dagger$ creates an electron in a state labeled by the band index $\lambda=+1$ and wave vector $\boldsymbol{k}$. The quantity $|0\rangle$ represents the vacuum state with no electrons in the conduction band.

By using the eigenstates in Eq. (4) of the main text, we find that

$$\langle\hat{n}_{q_z}\rangle_{\rm GS} = \langle\hat{n}_{q_z}\rangle_{\rm GS}^{(\rm scl)} + \langle\hat{n}_{q_z}\rangle_{\rm GS}^{(\rm int)}, \tag{D2}$$

where

$$\langle \hat{n}_{q_z} \rangle_{\rm GS}^{\rm (scl)} \equiv 2S \int_{k<k_{\rm F},k_z>0} \frac{d^3\boldsymbol{k}}{(2\pi)^3} \int dz \times e^{-iq_z z}\Theta(-z)[|u_{\boldsymbol{k},+1}|^2 + |u_{\bar{\boldsymbol{k}},+1}|^2] \tag{D3}$$

and

$$\langle \hat{n}_{q_z} \rangle_{\rm GS}^{\rm (int)} \equiv -2S \int_{k<k_{\rm F},k_z>0} \frac{d^3\boldsymbol{k}}{(2\pi)^3} \int dz \times e^{-iq_z z}\Theta(-z)[u^{\dagger}_{\bar{\boldsymbol{k}},+1}u_{\boldsymbol{k},+1}e^{i2k_z z} + u^{\dagger}_{\boldsymbol{k},+1}u_{\bar{\boldsymbol{k}},+1}e^{-i2k_z z}], \tag{D4}$$

where $u_{\boldsymbol{k},\lambda=\pm1}$ has been introduced in Eq. (2) of the main text. As in the main text, $\bar{\boldsymbol{k}} = (\boldsymbol{k}_\parallel, -k_z)$. The index "scl" ("int") stands for "semiclassical" ("interference"). Indeed, $\langle \hat{n}_{q_z} \rangle_{\rm GS}^{\rm (scl)}$ ($\langle \hat{n}_{q_z} \rangle_{\rm GS}^{\rm (int)}$) is independent of (dependent of) the relative phases—see Eq. (4) of the main text.

We start from the case $q_z = 0$. We find

$$\langle \hat{n}_0 \rangle_{\rm GS} = N + \langle \hat{n}_0 \rangle_{\rm GS}^{\rm (int)}, \tag{D5}$$

where N is the total number of electrons in the semi-infinite WSM, while the interference term $\langle \hat{n}_0 \rangle_{\rm GS}^{\rm int} \propto 1/L_z$ is negligible in the thermodynamic limit.

For a generic value of q_z, we obtain

$$\langle \hat{n}_{q_z} \rangle_{\rm GS}^{\rm (scl)} = nS\left[-\frac{1}{iq_z} + \pi\delta(q_z)\right], \tag{D6}$$

where $n \equiv 2N/(L_z S)$. We then calculate the interference term $\langle \hat{n}_{q_z} \rangle_{\rm GS}^{\rm (int)}$. We are interested in the small q_z limit. We find $\langle \hat{n}_{q_z} \rangle_{\rm GS}^{\rm (int)} - \langle \hat{n}_0 \rangle_{\rm GS}^{\rm (int)} \propto q_z$. In the thermodynamic limit, this can induce in $\mathrm{Re}[\chi^{(0)}_{\rm BB}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)]$ terms of $O(q_z^3,q_z'^3)$, which are beyond the interest of our leading-order long-wavelength theory.

Finally, taking into account only the semiclassical term in Eq. (D1), we find

$$\mathrm{Re}\big[\chi^{(0)}_{\rm BB}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)\big] = \frac{n\big(q_\parallel^2 + q_z q_z'\big)}{m_{\rm eff}L_z\omega^2} \times \left[\frac{1}{i(q_z' - q_z)} + \pi\delta(q_z - q_z')\right], \tag{D7}$$

which implies Eq. (14) of the main text.

APPENDIX E: THE AB RESPONSE FUNCTION

Here, we detail the calculation of the response function $\chi^{(0)}_{\rm AB}$, always for a generic doping. We neglect inter-Weyl-node electron-electron scattering processes and we linearize the spectrum $E_\lambda(\boldsymbol{k})$ around each Weyl node. In order to simplify the notation, we set $\hbar = 1$. Under these simplifying assumptions, the quantity we need to calculate is

$$\begin{aligned}\chi^{(0)}_{\rm AB}(q_z,q_z',\boldsymbol{q}_\parallel,\omega) = &\sum_{j=\pm1,\lambda=\pm1} \int \frac{d^3\boldsymbol{k}}{L_z(2\pi)^3}\Theta(-jk_x + jq_x)\Theta(k_z)\frac{\Theta(k_F + q_y - k_y) - \Theta(k_\lambda - k)}{vk_y - vq_y - \lambda vk + \omega + i\eta}\mathcal{M}^j_\lambda(q_x,q_z,q_z') \\ &+ \sum_{j=\pm1,\lambda=\pm1} \int \frac{d^3\boldsymbol{k}}{L_z(2\pi)^3}\Theta(-jk_x - jq_x)\Theta(k_z)\frac{\Theta(k_F - q_y - k_y) - \Theta(k_\lambda - k)}{vk_y + vq_y - \lambda vk - \omega - i\eta}\mathcal{M}^j_\lambda(-q_x,-q_z,-q_z'),\end{aligned} \tag{E1}$$

where $\eta = 0^+$, $j = \pm1$ is a Weyl-node index, $\lambda = \pm$ is a conduction/valence band index, $k_{+1} = k_{\rm F}$, and $k_{-1} = \Lambda$. Here, Λ is an ultraviolet cutoff. We will take the limit $\Lambda \to \infty$ momentarily. The matrix element $\mathcal{M}^j_\lambda(q_x,q_z,q_z')$ reads as follows,

$$\begin{aligned}\mathcal{M}^j_\lambda(q_x,q_z,q_z') = &\frac{-j(k_x - q_x)}{L_z}[1 + \lambda\cos(\beta_k)]\frac{1}{k_x^2 + k_z^2}\left[\frac{jk_x - ik_z}{-j(k_x - q_x) + i(k_z - q_z)} - \frac{jk_x + ik_z}{-j(k_x - q_x) - i(k_z + q_z)}\right] \\ &\times\left[\frac{jk_x + ik_z}{-j(k_x - q_x) - i(k_z - q_z')} - \frac{jk_x - ik_z}{-j(k_x - q_x) + i(k_z + q_z')}\right].\end{aligned} \tag{E2}$$

We denote by the shorthand q the small quantity $q_z \sim q_z' \sim q_x$, and we expand the response function up to the quadratic order in q. Carrying out the sum over the index j, we find

$$\chi^{(0)}_{\rm AB}(q_z,q_z',\boldsymbol{q}_\parallel,\omega) = -\frac{q_x^2 + q_z q_z'}{L_z\pi^3}\sum_\lambda \int d^3\boldsymbol{k}\left\{\frac{\Theta(k_x)\Theta(k_z)[\Theta(k_{\rm F} - k_y) - \Theta(k_\lambda - k)]}{v^2(k_y - \lambda k)^2 - (\omega + i\eta)^2}\right\}\frac{2\lambda vk_x k_z^2}{k\big(k_x^2 + k_z^2\big)^2}. \tag{E3}$$

It is now useful to define the following dimensionless quantities, $\boldsymbol{p} \equiv \boldsymbol{k}/k_{\rm F}$ and $\tilde{\omega} \equiv \omega/E_{\rm F}$, and use cylindrical coordinates, $p_z = p_\perp\cos(\phi)$ and $p_y = p_\perp\sin(\phi)$. We perform the integral over ϕ and, exploiting the identity

$\lim_{\Lambda\to\infty}[\Theta(1-p_y)-\Theta(\Lambda/k_F-p)]=-\Theta(p_y-1)$, we obtain

$$\chi^{(0)}_{AB}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)=\frac{q_x^2+q_zq_z'}{3L_z\pi^3E_F}[\mathcal{J}_1(\omega)-\mathcal{J}_2(\omega)-\mathcal{J}_3(\omega)], \tag{E4}$$

where we have defined

$$\mathcal{J}_1(\omega)=\int_{-\infty}^{+\infty}dp_y\int_0^\infty dp_\perp\frac{\Theta(1-p)}{p}\left[\frac{2}{(p_y-p)^2-(\tilde\omega+i\eta)^2}\right], \tag{E5}$$

$$\mathcal{J}_2(\omega)=\int_{-\infty}^{+\infty}dp_y\int_0^\infty dp_\perp\frac{\Theta(1-p_y)}{p}\left[\frac{2}{(p_y-p)^2-(\tilde\omega+i\eta)^2}\right], \tag{E6}$$

$$\mathcal{J}_3(\omega)=\int_{-\infty}^{+\infty}dp_y\int_0^\infty dp_\perp\frac{\Theta(p_y-1)}{p}\left[\frac{2}{(p_y+p)^2-(\tilde\omega+i\eta)^2}\right]. \tag{E7}$$

We first analyze the real part of the response function. We start by studying the auxiliary function $\mathcal{J}_1(\omega)$. Introducing a new set of polar coordinates, $p_y=p\sin(\theta)$ and $p_\perp=p\cos(\theta)$, we find

$$\mathrm{Re}[\mathcal{J}_1(\omega)]=\mathrm{P}\int_{-\pi/2}^{+\pi/2}d\theta\int_0^1 dp\frac{2}{p^2[\sin(\theta)-1]^2-\tilde\omega^2}, \tag{E8}$$

where "P" denotes the Cauchy principal value. The following mathematical identity will be used below,

$$\mathrm{P}\int dp\frac{1}{p^2[1-\sin(\theta)]^2-\tilde\omega^2}=\frac{1}{2\tilde\omega[1-\sin(\theta)]}\log\left|\frac{1-\frac{[1-\sin(\theta)]p}{\tilde\omega}}{1+\frac{[1-\sin(\theta)]p}{\tilde\omega}}\right|. \tag{E9}$$

Introducing the auxiliary variable $t=1-\sin(\theta)$, we can write

$$\mathrm{Re}[\mathcal{J}_1(\omega)]=\int_0^2\frac{dt}{\sqrt{t(2-t)}}\frac{1}{\tilde\omega t}\log\left|\frac{\tilde\omega-t}{\tilde\omega+t}\right|. \tag{E10}$$

We now note that the integrand in the previous equation is peaked at $t\sim 0$ and $t\sim\tilde\omega$. Furthermore, for $\omega\ll 2E_F$, both peaks collapse at $t\sim 0$. We can therefore approximate (E10) as

$$\mathrm{Re}[\mathcal{J}_1(\omega\ll 2E_F)]\simeq\int_0^2\frac{dt}{\sqrt{2t}}\frac{1}{\tilde\omega t}\log\left|\frac{\tilde\omega-t}{\tilde\omega+t}\right|, \tag{E11}$$

which can be evaluated analytically, yielding

$$\int_0^2\frac{dt}{\sqrt{2t}}\frac{1}{\tilde\omega t}\log\left|\frac{\tilde\omega-t}{\tilde\omega+t}\right|=-\frac{\sqrt{\tilde\omega}\log\left(\frac{2-\tilde\omega}{2+\tilde\omega}\right)+2\sqrt{2}\tanh^{-1}\left(\sqrt{2/\tilde\omega}\right)+2\sqrt{2}\tanh^{-1}\left(\sqrt{\tilde\omega/2}\right)}{\tilde\omega^{3/2}}. \tag{E12}$$

A further asymptotic expansion in the limit $\omega\ll 2E_F$ yields

$$\mathrm{Re}[\mathcal{J}_1(\omega\ll 2E_F)]\simeq-\frac{\sqrt{2}\pi}{\tilde\omega^{3/2}}. \tag{E13}$$

Similarly, we find

$$\mathrm{Re}[\mathcal{J}_2(\omega\ll 2E_F)]\simeq-\frac{\sqrt{2}\pi}{\tilde\omega^{3/2}}+\frac{\pi}{\sqrt{2\tilde\omega}}, \tag{E14}$$

while, in the same limits, $\mathrm{Re}[\mathcal{J}_3(\omega\ll 2E_F)]$ is a regular function having a negligible contribution with respect to $\mathrm{Re}(\mathcal{J}_1)$ and $\mathrm{Re}(\mathcal{J}_2)$. Replacing Eqs. (E13) and (E14) in Eq. (E4), we finally obtain

$$\mathrm{Re}\left[\chi^{(0)}_{AB}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)\right]\to-\frac{q_x^2+q_zq_z'}{3L_z\pi^2E_F}\frac{1}{\sqrt{2\tilde\omega}}, \tag{E15}$$

which gives Eq. (14) of the main text.

In order to evaluate the imaginary part of the response function, we exploit the Dirac identity $1/[x^2-(\omega+i0^+)^2]=1/(x^2-\omega^2)+i\pi\delta(x^2-\omega^2)$. For $\omega<2E_F$, we obtain

$$\mathrm{Im}[\mathcal{J}_1(\omega)]=\int dp_y\int_0^\infty dp_\perp\Theta(1-p)\delta(p_y-p+\tilde\omega)\frac{\pi}{\tilde\omega p}, \tag{E16}$$

$$\mathrm{Im}[\mathcal{J}_2(\omega)]=\int dp_y\int_0^\infty dp_\perp\Theta(1-p_y)\delta(p_y-p+\tilde\omega)\frac{\pi}{\tilde\omega p}, \tag{E17}$$

$$\mathrm{Im}[\mathcal{J}_3(\omega)]=0, \tag{E18}$$

which, replaced in Eq. (E4), result into

$$\mathrm{Im}\left[\chi^{(0)}_{\mathrm{AB}}(q_z,q_z',\boldsymbol{q}_\parallel,\omega)\right] = -\frac{q_x^2+q_zq_z'}{3L_z\pi^2E_{\mathrm{F}}}\frac{\left(\sqrt{2+\tilde{\omega}}-\sqrt{2-\tilde{\omega}}\right)}{\tilde{\omega}^{\frac{3}{2}}}. \tag{E19}$$

The leading term for $\omega \ll E_{\mathrm{F}}$ gives Eq. (23) of the main text.

Following similar steps in the undoped case ($E_{\mathrm{F}} = 0$), we find

$$\chi^{(0)}_{\mathrm{AB}}(q_z,q_z',\boldsymbol{q}_\parallel,\omega) \to -i\frac{2\left(q_x^2+q_zq_z'\right)}{3L_z\pi^2\omega}. \tag{E20}$$

[1] D. Pines and P. Noziéres, *The Theory of Quantum Liquids* (W. A. Benjamin, New York, 1966).

[2] P. M. Platzman and P. A. Wolff, *Waves and Interactions in Solid State Plasmas* (Academic, New York, 1973).

[3] G. F. Giuliani and G. Vignale, *Quantum Theory of the Electron Liquid* (Cambridge University Press, Cambridge, UK, 2005).

[4] S. A. Maier, *Plasmonics – Fundamentals and Applications* (Springer, New York, 2007).

[5] P. Alonso-González, A. Y. Nikitin, Y. Gao, A. Woessner, M. B. Lundeberg, A. Principi, N. Forcellini, W. Yan, S. Vélez, A. J. Huber, K. Watanabe, T. Taniguchi, F. Casanova, L. E. Hueso, M. Polini, J. Hone, F. H. L. Koppens, and R. Hillenbrand, Nat. Nanotechnol. **12**, 31 (2017).

[6] J. B. Khurgin, Nat. Nanotechnol. **10**, 2 (2015).

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

[8] 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 (2014).

[9] J. C. W. Song and M. S. Rudner, Proc. Natl. Acad. Sci. USA **113**, 4658 (2016).

[10] A. Kumar, A. Nemilentsau, K. H. Fung, G. Hanson, N. X. Fang, and T. Low, Phys. Rev. B **93**, 041413(R) (2016).

[11] P. Di Pietro, M. Ortolani, O. Limaj, A. Di Gaspare, V. Giliberti, F. Giorgianni, M. Brahlek, N. Bansal, N. Koirala, S. Oh, P. Calvani, and S. Lupi, Nat. Nanotechnol. **8**, 556 (2013).

[12] M. Autore, F. D'Apuzzo, A. Di Gaspare, V. Giliberti, O. Limaj, P. Roy, M. Brahlek, N. Koirala, S. Oh, F. J. G. de Abajo, and S. Lupi, Adv. Opt. Mater. **3**, 1257 (2015).

[13] N. Kumada, P. Roulleau, B. Roche, M. Hashisaka, H. Hibino, I. Petković, and D. C. Glattli, Phys. Rev. Lett. **113**, 266601 (2014).

[14] D. Jin, L. Lu, Z. Wang, C. Fang, J. D. Joannopoulos, M. Soljačić, L. Fu, and N. X. Fang, Nat. Commun. **7**, 13486 (2016).

[15] P. Hosur and X. Qi, C. R. Phys. **14**, 857 (2013).

[16] E. V. Gorbar, V. A. Miransky, I. A. Shovkovy, and P. O. Sukhachov, Phys. Rev. B **93**, 235127 (2016).

[17] M. Z. Hasan, S.-Y. Xu, I. Belopolski, and S.-M. Huang, Annu. Rev. Condens. Matter Phys. **8**, 289 (2017).

[18] B. Yan and C. Felser, Annu. Rev. Condens. Matter Phys. **8**, 337 (2017).

[19] A. A. Burkov, Annu. Rev. Condens. Matter Phys. **9**, 359 (2018).

[20] N. P. Armitage, E. J. Mele, and A. Vishwanath, Rev. Mod. Phys. **90**, 015001 (2018).

[21] C. Shekhar, A. K. Nayak, S. Singh, N. Kumar, S.-C. Wu, Y. Zhang, A. C. Komarek, E. Kampert, Y. Skourski, J. Wosnitza, W. Schnelle, A. McCollam, U. Zeitler, J. Kubler, S. S. P. Parkin, B. Yan, and C. Felser, arXiv:1604.01641.

[22] S.-Y. Xu, I. Belopolski, N. Alidoust, M. Neupane, G. Bian, C. Zhang, R. Sankar, G. Chang, Z. Yuan, C.-C. Lee, S.-M. Huang, H. Zheng, J. Ma, D. S. Sanchez, B. Wang, A. Bansil, F. Chou, P. P. Shibayev, H. Lin, S. Jia, and M. Z. Hasan, Science **349**, 613 (2015).

[23] B. Q. Lv, H. M. Weng, B. B. Fu, X. P. Wang, H. Miao, J. Ma, P. Richard, X. C. Huang, L. X. Zhao, G. F. Chen, Z. Fang, X. Dai, T. Qian, and H. Ding, Phys. Rev. X **5**, 031013 (2015).

[24] D. N. Basov, M. M. Fogler, and F. J. G. de Abajo, Science **354**, aag1992 (2016).

[25] F. M. D. Pellegrino, M. I. Katsnelson, and M. Polini, Phys. Rev. B **92**, 201407(R) (2015).

[26] J. Hofmann and S. D. Sarma, Phys. Rev. B **93**, 241402(R) (2016).

[27] F. Wilczek, Phys. Rev. Lett. **58**, 1799 (1987).

[28] J. C. W. Song and M. S. Rudner, Phys. Rev. B **96**, 205443 (2017).

[29] R. Okugawa and S. Murakami, Phys. Rev. B **89**, 235315 (2014).

[30] S. Rao, arXiv:1603.02821.

[31] J. M. Pitarke, V. M. Silkin, E. V. Chulkov, and P. M. Echenique, Rep. Prog. Phys. **70**, 1 (2007).

[32] I. Torre, M. I. Katsnelson, A. Diaspro, V. Pellegrini, and M. Polini, Phys. Rev. B **96**, 035433 (2017).

[33] R. H. Ritchie, Phys. Rev. **106**, 874 (1957).

[34] R. F. Wallis, J. J. Brion, E. Burstein, and A. Hartstein, Phys. Rev. B **9**, 3424 (1974).

[35] L. M Garrido, F. Flores and F. Garcia-Moliner, J. Phys. F: Met. Phys. **9**, 1097 (1979).

[36] J. Zhou, H. R. Chang, and D. Xiao, Phys. Rev. B **91**, 035114 (2015).

[37] M. Lv and S.-C. Zhang, Int. J. Mod. Phys. B **27**, 1350177 (2013).

[38] S. Tchoumakov, M. Civelli, and M. O. Goerbig, Phys. Rev. B **95**, 125306 (2017).

Correction: The ERC grant number contained an error and has been fixed, and a missing Mineco grant number has been inserted.