Inhomogeneous and nonlocal optical response in magic-angle twisted trilayer graphene

Chao Ding, Yueheng Du, and Mingwen Zhao [*]

School of Physics, Shandong University, Jinan 250100, China

The response of traditional two-dimensional (2D) materials to electromagnetic fields is typically homogeneous and local, as the wavelength of the electromagnetic field is much longer than the lattice constant of 2D materials. In contrast, plasmons in moiré flatband systems enable the coupling of electromagnetic fields and electrons at extremely small length scales. In this paper, we demonstrate that the plasmons in magic-angle twisted trilayer graphene (TTG) exhibit significant inhomogeneity and nonlocality, along with violation of in-plane wave-vector conservation. The exceptional optical scenarios arising from inhomogeneous response can be characterized by the 2D distribution patterns of the plasmonic electric fields. Our findings provide additional means for detecting the electronic states of materials by near-field optical experiments, as well as an innovative approach for advancing the development of optoelectronic devices.

I. INTRODUCTION

The response of two-dimensional (2D) materials to electromagnetic fields serves a valuable foundation for unveiling their inherent electronic structures, paving the way for various applications in photonics and optoelectronics, such as light-emitting devices, photodetectors, and ultrafast lasers [1–7]. In the realm of linear response, the current $J_\alpha(\mathbf{r}, \omega)$ $(\alpha = x, y)$ has a linear relationship with the applied electric field $E_\beta(\mathbf{r}', \omega)(\beta = x, y)$, connected by the conductivity $\sigma_{\alpha\beta}(\mathbf{r}, \mathbf{r}', \omega)$. In homogeneous response, such as that observed in systems with translation invariance, $\sigma_{\alpha\beta}(\mathbf{r}, \mathbf{r}', \omega)$ depends on $\mathbf{r} - \mathbf{r}'$, rather than $\mathbf{r}$ or $\mathbf{r}'$ [8]. When the current $J_\alpha(\mathbf{r}, \omega)$ relies solely on the electric field applied at the same position ($\mathbf{r}' = \mathbf{r}$), the response is termed localized. Otherwise, the response is referred to as inhomogeneous or nonlocalized. Typically, the interaction between conventional 2D materials and electromagnetic waves can be described as homogeneous and localized, because the wavelength of the electromagnetic wave significantly surpasses the lattice constant of these materials. This condition gives rise to position-independent optical conductivity $\sigma_{\alpha\beta}(\omega)$ $(\alpha, \beta = x, y)$. However, when the wavelength of an electromagnetic wave is comparable to the lattice constant of 2D materials, inhomogeneous optical response can become pronounced, leading to position-dependent optical conductivity [8]. Additionally, certain quantum effects, such as quantum pressure and diffusion, induce nonlocality to the response [9]. In contrast to the conventional optical response, inhomogeneous and nonlocal response provides more intricate insights into the electron structures of materials. Consequently, it offers an enhanced platform for exploring the short-range interactions within electron systems, including correlation effects [10].

As collective excitations of electrons, plasmons in 2D materials exhibit an extraordinary ability to confine optical fields on scales significantly shorter than wavelength of illumination [11–13], enabling a wide range of applications including nanophotonics, photodetection, and biosensing [14–18]. Among these materials, graphene plasmons stand out due to their exceptional properties, such as long lifetime, substantial wave localization, and high tunability [1,19,20]. The pronounced confinement of optical fields by graphene plasmons has been leveraged to explore nonlocal response effects, revealing that nonlocal response becomes appreciable in the phase space (q, ω) where ω/q approaches the Fermi velocity of graphene [10]. In the context of graphene superlattice, nonlocal response induces a decay channel for plasmons, resulting in substantial dissipative effects at low energies [21].

Twisted bilayer graphene (TBG) introduces an additional dimension for finely tuning its electronic structure by manipulating the twist angle. When the twist angle nears the magic angle $1.05°$, the energy bands near Fermi energy exhibit exceptional flatness, and are separated from other bands. This unique characteristic gives rise to the emergence of intriguing electronic phases, including correlated insulation and superconductivity [22–30]. The plasmon properties of TBG have garnered substantial attention from both theoretical and experimental studies [31–39]. It has been demonstrated that magic-angle TBG is capable of sustaining undamped plasmon modes, distinguished by their remarkable longevity and significant wave vectors extending to the boundaries of the Brillouin zone (BZ) [31]. This undamped property of the moiré flat plasmons (MFPs) can be attributed to the relatively large values of the "fine structure" constant $\alpha = e^2/\hbar\kappa v_F$ arising from the low Fermi velocity v_F of moiré flatbands, which raises the plasmon energies beyond the realm of single-particle excitation [31]. Significantly, the wavelength of these undamped MFPs can be reduced to the order of 20–30 nm, a scale comparable to the lattice constants of magic-angle TBG [40], hinting at a substantial inhomogeneous optical response within the material.

Moiré flatbands also manifest in twisted trilayer graphene (TTG) systems when the magic angle of about $1.6°$ is achieved

*zmw@sdu.edu.cn

[41–43]. The larger magic angle in TTG, which is about $\sqrt{2}$ times larger than that of TBG, can be attributed to the strengthened effective interlayer coupling [41–43]. Similar to TBG, magic-angle TTG also supports MFPs [44]. The enhanced interlayer coupling in TTG compared to TBG may lead to more pronounced inhomogeneous response. Furthermore, the increased magic angle results in a shorter moiré lattice constant (∼8.8 nm) for TTG, rendering it more suitable for investigating response properties at smaller length scales. In this paper, we delve into the role of inhomogeneous and nonlocal effects on the plasmonic properties of magic-angle TTG. Our findings reveal that the position-dependent interlayer coupling, which breaks the transition invariant within TTG, leads to a pronounced inhomogeneous response. This inhomogeneous response can be identified by a 2D periodic distribution of plasmon electric fields in real space, which differs significantly from the one-dimensional patterns observed in a homogeneous response. This intriguing scenario unveils additional avenues for exploring the electronic states of materials through near-field optical experiments. Moreover, we demonstrate that the TTG exhibits remarkable nonlocal response characteristics stemming from quantum origins.

II. THEORETICAL APPROACH

A. General theory for the optical response in 2D electron systems

In the framework of linear response theory, the general formula of current $J_\alpha(\mathbf{r}, \omega)$ in response to the electric field $E_\beta(\mathbf{r}', \omega)$ can be written as [33]

$$J_\alpha(\mathbf{r}, \omega) = \sum_\beta \int \sigma_{\alpha\beta}(\mathbf{r}, \mathbf{r}', \omega)E_\beta(\mathbf{r}', \omega)d\mathbf{r}', \qquad (1)$$

where $\sigma_{\alpha\beta}(\mathbf{r}, \mathbf{r}', \omega)$ denotes the component of the optical conductivity tensor. The Fourier transform with respect to $\mathbf{r}$ and $\mathbf{r}'$ leads to the expression

$$J_\alpha(\mathbf{q}, \omega) = \sum_{\beta, \mathbf{q}'} \sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega)E_\beta(\mathbf{q}', \omega), \qquad (2)$$

with

$$\sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega) = \frac{1}{S} \iint \sigma_{\alpha\beta}(\mathbf{r}, \mathbf{r}', \omega)e^{-i\mathbf{q}\cdot\mathbf{r}}e^{i\mathbf{q}'\cdot\mathbf{r}'}d\mathbf{r}d\mathbf{r}', \quad (3)$$

and S is the area of the 2D system.

In the homogeneous response limit, e.g., in the system with translation invariance, the optical conductivity $\sigma_{\alpha\beta}(\mathbf{r}, \mathbf{r}', \omega)$ is solely determined by the vector $\mathbf{r} - \mathbf{r}'$ [8]. Consequently, the conductivity tensor reduces to $\sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega) = \sigma_{\alpha\beta}(\mathbf{q}, \omega)\delta_{\mathbf{q},\mathbf{q}'}$, corresponding to the conservation of momentum ($\mathbf{q} = \mathbf{q}'$) during optical response. However, for the inhomogeneous response, there exists $\sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega) \neq 0$ for $\mathbf{q} \neq \mathbf{q}'$, implying the broken momentum conservation.

In the local response limit, the current at a given point $\mathbf{r}$ is exclusively determined by the electric field at the same location. Consequently, we derive the optical conductivity tensor for this scenario, $\sigma_{\alpha\beta}(\mathbf{r}, \mathbf{r}', \omega) = \sigma_{\alpha\beta}(\mathbf{r}, \omega)\delta(\mathbf{r} - \mathbf{r}')$, which leads to $\sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega) = \sigma_{\alpha\beta}(\mathbf{q} - \mathbf{q}', 0, \omega)$ according to Eq. (3). Therefore, within the framework of homogeneity and the local response limit, we derive an optical

conductivity tensor, $\sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega) = \sigma_{\alpha\beta}(\mathbf{q} \to 0, \omega)\delta_{\mathbf{q},\mathbf{q}'} \equiv \sigma_{\alpha\beta}(\omega)\delta_{\mathbf{q},\mathbf{q}'}$. The q-independent optical conductivity tensor $\sigma_{\alpha\beta}(\omega)$ was extensively utilized in prior studies.

B. Dielectric function and plasmons in 2D electron systems

When an external scalar potential energy $V_{\text{ext}}(\mathbf{r}, \omega)$ is applied to a 2D electron system, the redistribution of electrons leads to an additional potential $V_{\text{ind}}(\mathbf{r}, \omega)$, with

$$V_{\text{ind}}(\mathbf{r}, \omega) = \frac{e^2}{4\pi\varepsilon_0} \int \frac{\Delta n(\mathbf{r}', \omega)}{|\mathbf{r} - \mathbf{r}'|}d\mathbf{r}', \qquad (4)$$

where $\Delta n(\mathbf{r}', \omega)$ is the fluctuation of electron density with respect to that of ground state. Notably, $V_{\text{ind}}(\mathbf{r}, \omega)$ differs from the Hartree potentials arising from the charge inhomogeneity of ground state [33].

Total potential energy is given as

$$V_{\text{tot}}(\mathbf{r}, \omega) = V_{\text{ext}}(\mathbf{r}, \omega) + V_{\text{ind}}(\mathbf{r}, \omega), \qquad (5)$$

According to the linear response theory, $\Delta n(\mathbf{r}, \omega)$ is correlated to $V_{\text{tot}}(\mathbf{r}', \omega)$ through a density-density response function $\chi(\mathbf{r}, \mathbf{r}', \omega)$ [8]:

$$\Delta n(\mathbf{r}, \omega) = \int \chi(\mathbf{r}, \mathbf{r}', \omega)V_{\text{tot}}(\mathbf{r}', \omega)d\mathbf{r}'. \qquad (6)$$

The combination of Eq. (4)–Eq. (6) allows us to define a dielectric function $\varepsilon(\mathbf{r}, \mathbf{r}', \omega)$, serving as a connection between $V_{\text{ext}}(\mathbf{r}, \omega)$ and $V_{\text{tot}}(\mathbf{r}', \omega)$,

$$V_{\text{ext}}(\mathbf{r}, \omega) = \int \varepsilon(\mathbf{r}, \mathbf{r}', \omega)V_{\text{tot}}(\mathbf{r}', \omega)d\mathbf{r}'. \qquad (7)$$

Alternatively, dielectric function can also be expressed in wave-vector space through a Fourier transform,

$$V_{\text{ext}}(\mathbf{q}, \omega) = \sum_{\mathbf{q}'} \varepsilon(\mathbf{q}, \mathbf{q}', \omega)V_{\text{tot}}(\mathbf{q}', \omega), \qquad (8)$$

with

$$\varepsilon(\mathbf{q}, \mathbf{q}', \omega) = \delta_{\mathbf{q},\mathbf{q}'} - v(\mathbf{q})\chi(\mathbf{q}, \mathbf{q}', \omega) \qquad (9)$$

where $v(\mathbf{q}) = e^2/(2\varepsilon_0\varepsilon_r q)$ is the Fourier transform of 2D Coulomb potential, and ε_r is the background dielectric constant. The density-density response function $\chi(\mathbf{q}, \mathbf{q}', \omega)$ can be calculated within the random-phase approximation,

$$\chi(\mathbf{q}, \mathbf{q}', \omega) = \frac{1}{S} \sum_{\alpha,\beta} \frac{f(E_\alpha) - f(E_\beta)}{\hbar\omega - E_\beta + E_\alpha + i\eta}F_{\alpha\beta}(\mathbf{q}, \mathbf{q}'), \quad (10)$$

where S is the area of the 2D system, α and β represent the eigenstates states with the energy of E_α and E_β, $f(E)$ is the Fermi distribution function, and $F_{\alpha\beta}(\mathbf{q}, \mathbf{q}') = \langle\alpha|e^{-i\mathbf{q}\cdot\mathbf{r}}|\beta\rangle\langle\beta|e^{i\mathbf{q}'\cdot\mathbf{r}}|\alpha\rangle$.

Importantly, using the charge-continuity equation $\nabla \cdot \mathbf{J}(\mathbf{r}, \omega) = -i\omega e\Delta n(\mathbf{r}, \omega)$ and $\mathbf{E}(\mathbf{r}, \omega) = \frac{1}{e}\nabla V_{\text{tot}}(\mathbf{r}, \omega)$, along with Eqs. (1) and (6), we can establish the relation between the density-density response function and conductivity,

$$\chi(\mathbf{q}, \mathbf{q}', \omega) = \frac{1}{ie^2\omega} \sum_{\alpha,\beta} q_\alpha q'_\beta \sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega). \qquad (11)$$
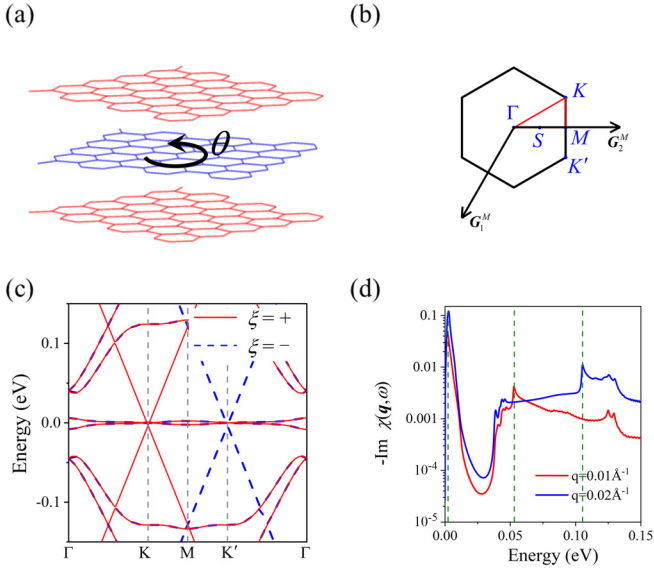
FIG. 1. (a) Schematic of TTG, where the top and bottom layers are aligned, whereas the middle layer is relatively twisted with θ. (b) The moiré BZ of TTG. $\mathbf{G}_1^M$ and $\mathbf{G}_2^M$ represent the reciprocal lattice vectors. (c) The band structures of TTG at magic angle $\theta = 1.6°$. The bands for valley $\xi = +$ and $\xi = -$ are represented by red solid lines and blue dashed lines, respectively. (d) $-\text{Im}\,\chi(\mathbf{q}, \omega)$ for $\mathbf{q}$ along x axis, with $q = 0.01$ Å^{-1} or $q = 0.02$ Å^{-1}, respectively.

Furthermore, in the framework of homogeneous and the local response limit, we can obtain

$$\chi(\mathbf{q}, \omega) \approx \frac{1}{ie^2\omega} \sum_{\alpha,\beta} q_\alpha q_\beta \sigma_{\alpha\beta}(\omega). \tag{12}$$

For the optical response in 2D crystals, the lattice periodicity leads to the quasimomentum conservation $\mathbf{q}' = \mathbf{q} + \mathbf{G}$ in the optical response, where $\mathbf{G}$ is the reciprocal lattice, due to the lattice periodicity. Consequently, the dielectric function becomes a matrix defined as $\varepsilon_{\mathbf{G},\mathbf{G}'}(\mathbf{q}, \omega) = \varepsilon(\mathbf{q} + \mathbf{G}, \mathbf{q} + \mathbf{G}', \omega)$, and the plasmon modes are determined by $\det \varepsilon_{\mathbf{G},\mathbf{G}'}(\mathbf{q}, \omega) = 0$. This condition is reduced to $\varepsilon_{0,0}(\mathbf{q}, \omega) = 0$ in the homogeneous response limit, owing to the momentum conservation $\mathbf{q}' = \mathbf{q}$ in the response process.

III. RESULTS AND DISCUSSION

A. Electronic band structures and polarization function of TTG

We consider a specific TTG to demonstrate the inhomogeneous and nonlocal response effect. The TTG is constructed by rotating the middle layer of an AAA stacked trilayer graphene by an angle of θ with respect to the two outer layers, as depicted in Fig. 1(a) [41]. The electronic band structure of the TTG can be calculated by using a continuum model. In this model, the Hamiltonian is constructed by the low-energy electronic states near the Dirac points $K^{(+)}$ or $K^{(-)}$ of graphene. The Hamiltonian for valley $K^{(\xi)}$ is given as

$$H^{(\xi)} = \begin{pmatrix} H_1 & T^+ & 0 \\ T & H_2 & T \\ 0 & T^+ & H_3 \end{pmatrix}, \tag{13}$$

where $H_l = -\hbar v\{R[(-1)^{l+1}\theta/2](\mathbf{k} - K_l^{(\xi)})\} \cdot (\xi\sigma_x, \sigma_y)$ represents the intralayer contribution of layer l. $R(\varphi)$ is a 2×2 rotation matrix, and $K_l^{(\xi)}$ denotes the Dirac point for layer l at valley $K^{(\xi)}$. The interlayer coupling is characterized by a moiré potential T, which is given by

$$T(\mathbf{r}) = \begin{pmatrix} u & u' \\ u' & u \end{pmatrix} + e^{i\xi\mathbf{G}_1^M \cdot r}\begin{pmatrix} u & u'\omega^{-\xi} \\ u'\omega^\xi & u \end{pmatrix}$$
$$+ e^{i\xi\left(\mathbf{G}_1^M + \mathbf{G}_2^M\right)\cdot r}\begin{pmatrix} u & u'\omega^\xi \\ u'\omega^{-\xi} & u \end{pmatrix}. \tag{14}$$

In this expression, $\mathbf{G}_1^M$ and $\mathbf{G}_2^M$ are the moiré reciprocal lattice vectors, as shown in Fig. 1(b). In our calculations, we take $u = 79.7$meV and $u' = 97.5$meV as adopted in previous literature, which are obtained by a tight-binding model with intra- and interlayer hopping amplitudes given by $t_0 = -2.7$eV and $t_\perp = 0.48$eV, respectively [23]. The difference between u and u' is due to the interlayer corrugation effect. The interlayer coupling between two outer layers was not taken into account, due to its negligible magnitude (< 7meV) [41].

The Hamiltonian of Eq. (13) can be rewritten in the bonding–antibonding bases as [41]

$$\tilde{H}^{(\xi)} = \begin{pmatrix} H_+ & \sqrt{2}T & 0 \\ \sqrt{2}T^+ & H_2 & 0 \\ 0 & 0 & H_- \end{pmatrix}, \tag{15}$$

where $H_{+/-}$ represents the Hamiltonian of top and bottom layers, respectively. Importantly, we can decompose $\tilde{H}^{(\xi)}$ into a bilayer Hamiltonian that closely resembles the TBG Hamiltonian but with a proportionally increased interlayer coupling by a factor of $\sqrt{2}$ factor and that of a single-layer Hamiltonian of graphene. As a result, the band structure of TTG combines that of TBG with a significantly enhanced interlayer coupling and that of pristine graphene. The first magic angle in TTG is enlarged by a factor of $\sqrt{2}$ compared to TBG, due to the increased effective interlayer coupling, which is about $\theta = 1.6°$ observed in experimental results [42].

The electronic band structure near the Fermi level (set to zero) of TTG at the magic angle of $\theta = 1.6°$ obtained from the above Hamiltonian is depicted in Fig. 1(c). In the $\xi = +$ valley, we observe the presence of two moiré flatbands and a Dirac cone (DC) with Dirac point located at K, which can be attributed to a combination of the bilayer Hamiltonian and a single-layer Hamiltonian of graphene as expressed in Eq. (15). The Dirac point is found to be slightly below the Fermi level by about 3.4 meV. The bands for the $\xi = -$ valley mirror those of the $\xi = +$ valley due to time-reversal symmetry.

Notably, for $\hbar\omega \gg t_\perp$, the interlayer-coupling effect is neglectable, and the TTG can thus be treated as three decoupled graphene monolayers. The polarization function becomes

$$\chi(\mathbf{q}, \omega) \approx \frac{-ig_l g_v g_s}{16\hbar} \frac{q^2}{\sqrt{\omega^2 - (v_F q)^2}}, \tag{16}$$

at zero temperature and zero chemical potential, where $g_l = 3$, $g_v = g_s = 2$ are the layer, valley, and spin degeneracy, respectively. v_F is the Fermi velocity of graphene monolayer. This result is similar to that of TBG [32].
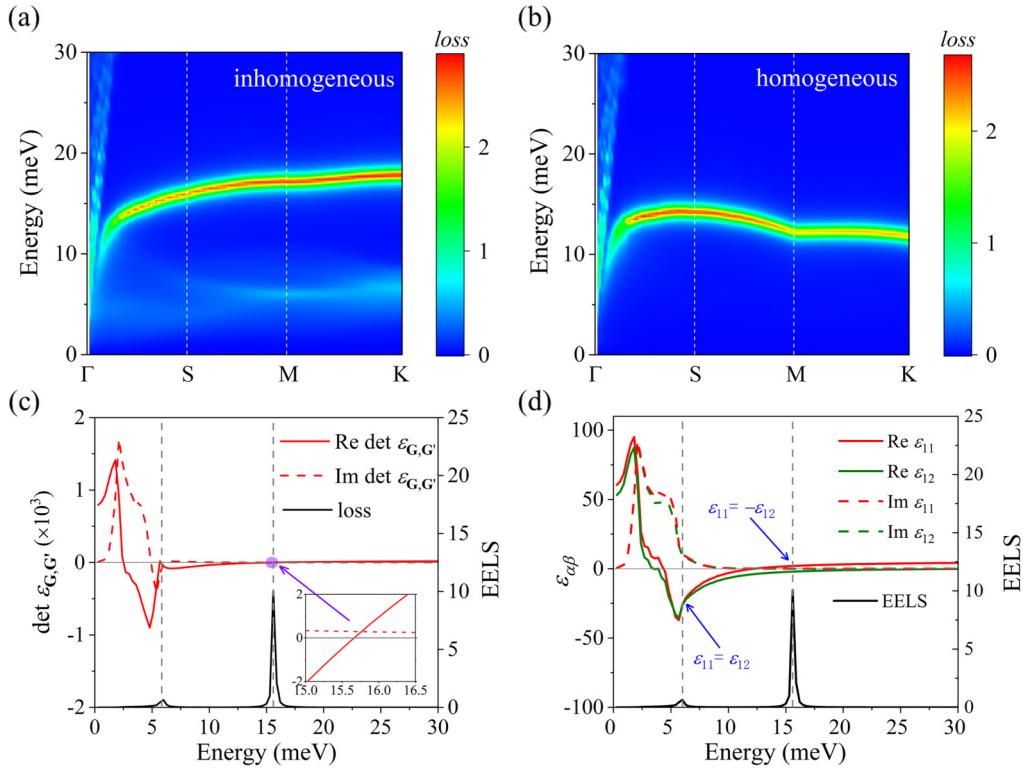
FIG. 2. (a) The EELS as the functions of wave vector $\mathbf{q}$ (along the high-symmetry lines) and energy $\hbar\omega$. The results for homogeneous response limit are shown in (b). (c) The real and imaginary parts of det $\varepsilon_{\mathbf{G},\mathbf{G}'}$, and EELS for $\mathbf{q} = M$, calculated by a 2×2 dielectric function matrix. The undamped plasmon mode corresponds to the peak value of loss function at $\hbar\omega \approx 15.6$ meV, where Re det $\varepsilon_{\mathbf{G},\mathbf{G}'} = 0$, as shown in the illustration. (d) The real and imaginary parts of dielectric function matrix elements ε_{11} and ε_{12}, as well as EELS for $\mathbf{q} = M$. For the damped plasmon mode at lower energy, it has $\varepsilon_{11} = \varepsilon_{12}$, whereas they are opposite for undamped plasmon mode at higher energy.

However, when $\hbar\omega \ll t_\perp$, the interlayer coupling plays an important rule. The interlayer coupling leads to flatbands with the renormalized Fermi velocity $v_F^* = v_F(1-6\alpha^2)/(1 + 12\alpha^2)$, where $\alpha = u'/2k_D v_F \sin(\theta/2)$ is a dimensionless parameter characterizing the interlayer coupling [41]. Therefore, we have

$$\chi(\mathbf{q},\omega) \approx \frac{-ig_v g_s}{16\hbar}(q^2/\sqrt{\omega^2-(v_F q)^2} + 2q^2/\sqrt{\omega^2-(v_F^* q)^2}). \tag{17}$$

The polarization function diverges when $\hbar\omega \approx \hbar v_F q$ or $\hbar\omega \approx \hbar v_F^* q$. This contrasts with the polarization function of TBG, which diverges only at $\hbar\omega = \hbar v_F^* q$ [32]. This result was also verified from our numerical calculations, as depicted in Fig. 1(d). The two peaks of $-\mathrm{Im}\chi(\mathbf{q},\omega)$ for $q_x = 0.01$ and $0.02\,\text{Å}^{-1}$ are contributed by moiré flatbands ($\hbar\omega \approx \hbar v_F^* q$) and graphene monolayer ($\hbar\omega \approx \hbar v_F q$), respectively.

B. Inhomogeneous response in TTG

Based on the electronic band structure of TTG, we explore the impact of inhomogeneous response in TTG's plasmonic behavior. The plasmon modes determined by det $\varepsilon_{\mathbf{G},\mathbf{G}'}(\mathbf{q},\omega) = 0$ enable the definition of the electron-energy-loss spectrum (EELS) as EELS = $-\mathrm{Im}[Tr\varepsilon_{\mathbf{G},\mathbf{G}'}^{-1}(\mathbf{q},\omega)]$ [40]. Identification of plasmon modes is achieved through the peak values of EELS, as illustrated in Fig. 2. For the inhomogeneous response depicted in

Fig. 2(a), we adopted 9 reciprocal lattice vectors, $\mathbf{G}, \mathbf{G}' = m\mathbf{G}_1^M + n\mathbf{G}_2^M$, with $m, n \in (-1, 0, 1)$, corresponding to 81 Umklapp processes. This choice ensures convergence of dielectric function matrix calculations [32]. The EELS in the homogeneous response limit was obtained by excluding the Umklapp processes, as shown in Fig. 2(b). In both scenarios, an undamped plasmon branch extending across the entire BZ is evident. These plasmon modes can be attributed to the electron transitions between the moiré flatbands, resulting in small energy fluctuation smaller than 10 meV. For small wave vectors $\mathbf{q}$, the disparity between the real plasmon dispersion and that under the homogeneous response limit is neglectable, implying a weak inhomogeneous response effect in the long-wavelength approximation. However, as the wave vector increases, the homogeneous response limit fails to reproduce the positive group velocities of plasmon dispersion, indicating an enhanced inhomogeneous response effect in TTG for large wave vectors. Furthermore, the inhomogeneous response leads to the emergency of damped plasmons in the frequency region around 5 meV.

To reveal the role of the inhomogeneous response in the MFPs in TTG, we constructed a minimal model of the dielectric function matrix $\varepsilon_{\mathbf{G},\mathbf{G}'}(\mathbf{q},\omega)$ at $\mathbf{q} = M$, focusing on a restricted set of reciprocal lattice vectors, $\mathbf{G}, \mathbf{G}' \in \{0, -\mathbf{G}_2^M\}$, as

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{pmatrix}, \tag{18}$$

with $\varepsilon_{11}=\varepsilon_{0,0}$, $\varepsilon_{22}=\varepsilon_{-\boldsymbol{G}_2^M,-\boldsymbol{G}_2^M}$, $\varepsilon_{12}=\varepsilon_{0,-\boldsymbol{G}_2^M}$, and $\varepsilon_{21}=\varepsilon_{-\boldsymbol{G}_2^M,0}$. The C_2 symmetry of TTG leads to $\varepsilon_{11}=\varepsilon_{22}$ and $\varepsilon_{12}=\varepsilon_{21}$, as demonstrated in Appendix A. The two plasmon modes given by $\det \varepsilon_{\boldsymbol{G},\boldsymbol{G}'}=0$ correspond to $\varepsilon_{11}=\pm\varepsilon_{12}$, respectively. We plotted the real and imaginary parts of $\det\varepsilon_{\boldsymbol{G},\boldsymbol{G}'}$, Re[$\det\varepsilon_{\boldsymbol{G},\boldsymbol{G}'}$], and Im[$\det\varepsilon_{\boldsymbol{G},\boldsymbol{G}'}$], along with the EELS of TTG in Fig. 2(c). The two zeros of Re[$\det\varepsilon_{\boldsymbol{G},\boldsymbol{G}'}$] at $\hbar\omega\approx 5.7$ and 15.6 meV are along with the two peaks of EELS at these energies, indicating the presence of plasmon modes at these frequencies. Notably, at the frequency of $\hbar\omega\approx 5.7$meV, Im[$\det\varepsilon_{\boldsymbol{G},\boldsymbol{G}'}$] has a finite value, indicating the presence of a damped plasmon mode. In contrast, for the higher-frequency mode, Im[$\det\varepsilon_{\boldsymbol{G},\boldsymbol{G}'}$] approaches nearly zero, signifying an undamped plasmon mode. We also depicted the spectra of ε_{11} and ε_{12} in Fig. 2(d). The lower-frequency damped plasmon mode and the higher-frequency undamped plasmon mode are consistent with the conditions $\varepsilon_{11}=\varepsilon_{12}$ and $\varepsilon_{11}=-\varepsilon_{12}$, respectively, as we have analyzed. In the case of homogeneous response limit, $\varepsilon_{\boldsymbol{G},\boldsymbol{G}'}(\boldsymbol{q},\omega)$ becomes a diagonal matrix, and the plasmon modes are determined by the zeros of the diagonal elements of ε, i.e., $\varepsilon_{11}=\varepsilon_{22}=0$, which gives two degenerate plasmon frequencies at $\hbar\omega\approx 12.3$meV. However, the nonzero ε_{12} term stemming from the inhomogeneous response leads to the splitting of two modes, resulting an undamped plasmon mode of larger frequency $\hbar\omega\approx 15.6$meV and a damped plasmon mode of small frequency $\hbar\omega\approx 5.7$ meV.

The inhomogeneous response of TTG can be correlated to the Umklapp processes. The periodic interlayer moiré potential breaks the translational symmetry of the TTG, allowing the emergence of $\sigma_{\alpha\beta}(\boldsymbol{q},\boldsymbol{q}',\omega)\neq 0$ at $\boldsymbol{q}'=\boldsymbol{q}+\boldsymbol{G}$, where $\boldsymbol{G}$ represents the reciprocal moiré lattice vector. For the moiré potential with a small periodicity, e.g., in the TTG with a large twisted angle, Umklapp processes are suppressed because the wave vector is significantly smaller than the length of the reciprocal moiré lattice vector. Consequently, the response can be treated homogeneously. Conversely, for the TTG with a small twisted angle close to the magic angle, leading to a larger periodicity, the wave vector becomes comparable to the length of the reciprocal moiré lattice vector, making the Umklapp processes significant. In such cases, the response exhibits pronounced inhomogeneous feature. This mechanism also holds for TBG, where interlayer moiré potentials play a crucial role. It has been reported that the existence of damped plasmon branch in magic-angle TBG is closely related to the inhomogeneous response [45].

C. Electric field distribution of MFPs

To visualize the inhomogeneous response features in TTG, we further determined the distribution of plasmon electric fields on the TTG. In our model system, TTG is placed at $z=0$ plane surrounded by a medium with dielectric constant ε and magnetic permeability $\mu=1$, as shown in Fig. 3(a). Generally, the incident electromagnetic wave for exciting plasmons can be expressed as

$$\mathbf{E}_i(\boldsymbol{q},\omega)=\sum_{\boldsymbol{G}}(E_{i,\mathbf{q}+\boldsymbol{G}}\mathbf{e}_{\mathbf{q}+\boldsymbol{G}}+E_{i,z}\mathbf{e}_z)e^{i(\mathbf{q}+\boldsymbol{G})\cdot\mathbf{r}}e^{-\kappa z}e^{-i\omega t}(\kappa > 0),$$
(19)

where $\mathbf{q}=(q_x,q_y)$ and $\mathbf{r}=(x,y)$ are, respectively, the wave vector and position vector projected onto the xy plane. $\mathbf{e}_{\mathbf{q}+\boldsymbol{G}}$ and $\mathbf{e}_z$ are the unit vectors parallel to $\boldsymbol{q}+\boldsymbol{G}$ and along the z direction, respectively. The wave equation from Maxwell's equations,$\nabla^2\mathbf{E}_i-\mu_0\varepsilon_0\varepsilon\partial^2\mathbf{E}_i/\partial t^2=0$, gives $\kappa=\sqrt{|\mathbf{q}+\mathbf{G}|^2-\varepsilon\omega^2/c^2}$. In the unretarded limit ($|\mathbf{q}+\mathbf{G}|\gg\sqrt{\varepsilon}\omega/c$), we have $\kappa\approx|\mathbf{q}+\mathbf{G}|$. In experiments, $\mathbf{E}_i$ can be obtained by Otto configuration, where the wave vector of light is increased by a prism of large dielectric constant [13,46,47].

Concerning the quasimomentum conservation ($\mathbf{q}'=\mathbf{q}+\boldsymbol{G}'$) in TTG, we assume the reflection wave $\mathbf{E}_r$ and transmission waves $\mathbf{E}_t$ as

$$\mathbf{E}_r(\boldsymbol{q},\omega)=\sum_{\boldsymbol{G}'}[E_{r,\mathbf{q}+\boldsymbol{G}'}\mathbf{e}_{\mathbf{q}+\boldsymbol{G}'}+E_{r,z}\mathbf{e}_z]e^{i(q+\boldsymbol{G}')\cdot\mathbf{r}}e^{|q+\boldsymbol{G}'|z}e^{-i\omega t},$$
(20)

and

$$\mathbf{E}_t(\boldsymbol{q},\omega)=\sum_{\boldsymbol{G}'}[E_{t,\mathbf{q}+\boldsymbol{G}'}\mathbf{e}_{\mathbf{q}+\boldsymbol{G}'}+E_{t,z}\mathbf{e}_z]e^{i(q+\boldsymbol{G}')\cdot\mathbf{r}}e^{-|q+\boldsymbol{G}'|z}e^{-i\omega t}.$$
(21)

All the components of reflection and transmission waves are evanescent waves due to $|\mathbf{q}+\boldsymbol{G}'|>\sqrt{\varepsilon}\omega/c$, which will decay to zero when $z\to -\infty$ (for $\mathbf{E}_r$) and $z\to +\infty$ (for $\mathbf{E}_t$). $\boldsymbol{E}_i(\boldsymbol{q},\omega)$ and $\boldsymbol{E}_t(\boldsymbol{q},\omega)$ can be represented by vector matrixes in the reciprocal space: $\mathrm{E}_{\mathbf{i}}=\{E_{i,\mathbf{q}+\boldsymbol{G}_1},E_{i,\mathbf{q}+\boldsymbol{G}_2},...\}$ and $\mathrm{E}_t=\{E_{t,\mathbf{q}+\boldsymbol{G}_1},E_{t,\mathbf{q}+\boldsymbol{G}_2},...\}$. According to the Maxwell electric field theory, we derive a simple equation that establishes a connection between $\mathbf{E_t}$ to $\mathbf{E_i}$, as follows (see Appendix B for details):

$$\mathcal{F}^{-1}\mathbf{E_t}=\mathbf{E_i},$$
(22)

with $\mathcal{F}^{-1}=\mathbf{I}-\chi\mathbf{V}$, where $\mathbf{I}$ is an identity matrix, χ is the density-density response function matrix defined as $\chi_{\boldsymbol{G},\boldsymbol{G}'}=\chi(\mathbf{q}+\boldsymbol{G},\mathbf{q}+\boldsymbol{G}',\omega)$, and $\mathbf{V}$ is a diagonal matrix of 2D Coulomb potential expressed as $V_{\boldsymbol{G},\boldsymbol{G}'}=e^2/(2\varepsilon_0\varepsilon|\mathbf{q}+\boldsymbol{G}|)\delta_{\boldsymbol{G},\boldsymbol{G}'}$. Clearly, in the case of homogeneous response, both and $\mathcal{F}^{-1}$ are diagonal, because χ is diagonal. Consequently, $E_{t,\mathbf{q}+\boldsymbol{G}_n}$ is nonzero only when the corresponding incident wave component $E_{i,\mathbf{q}+\boldsymbol{G}_n}$ is nonzero, and the wave vector is conserved in the in-plane direction ($\mathbf{q}'=\mathbf{q}$). For inhomogeneous responses, however, the nonzero off-diagonal components in can lead to nonzero $E_{t,\mathbf{q}+\boldsymbol{G}_n}$ even when $E_{i,\mathbf{q}+\boldsymbol{G}_n}=0$, breaking the in-plane wave-vector conservation.

To further verify the above analysis, we considered a monochromic electromagnetic plane wave, $\mathbf{E}_i(\boldsymbol{q},\omega)=(E_{i,\mathbf{q}}\mathbf{e}_{\mathbf{q}}+E_{i,z}\mathbf{e}_z)e^{i\mathbf{q}\cdot\mathbf{r}}e^{-qz}e^{-i\omega t}$ as incident wave and calculated the transmission coefficient $t(\mathbf{q}+\boldsymbol{G}')=E_{t,\mathbf{q}+\boldsymbol{G}'}/E_{i,q}$ as a function of frequency ω for $\mathbf{q}=S,M,$ and$K$, as depicted in Figs. 3(b)–3(d). For each incident wave vector $\mathbf{q}$, we checked nine transmission wave vectors $\mathbf{q}'=\mathbf{q}+\boldsymbol{G}'$, with $\boldsymbol{G}'\in 0\pm\boldsymbol{G}_1^M,\pm\boldsymbol{G}_2^M,\pm\boldsymbol{G}_1^M\pm\boldsymbol{G}_2^M$, labeled as $1\sim 9(\mathbf{q}'_1\sim\mathbf{q}'_9)$. From Fig. 3(b), for the case of $\mathbf{q}=$ S, we can observe that besides $-\mathrm{Im}[t(\mathbf{q})]$, $-\mathrm{Im}[t(\mathbf{q}'_5)]=-\mathrm{Im}[t(\mathbf{q}-\boldsymbol{G}_2^M)]$ has a substantial value, verifying the failure of in-plane wave-vector conservation in TTG. Similar results are also observed for $\mathbf{q}=M$ and K.

Notably, in the absence of incident electromagnetic waves, i.e.,$E_{i,q}=0$, the nontrivial solutions of Eq. (22) given by $\det\mathcal{F}^{-1}=0$ correspond to the plasmon modes of the
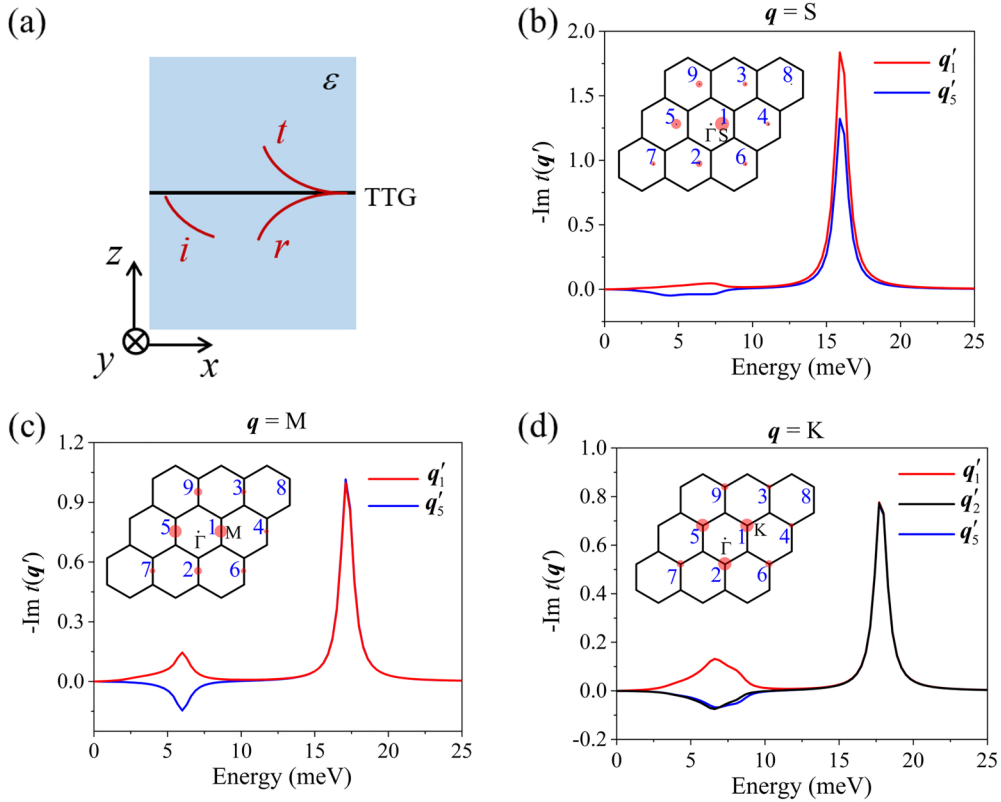
FIG. 3. (a) Schematic of the excitation of plasmons of TTG by using evanescent wave. TTG is placed at $z = 0$ plane surrounded by a medium with dielectric constant ε and magnetic permeability $\mu = 1$. i, r, and t represent the incident, reflected, and transmitted waves, respectively. (b)–(d) The $-\mathrm{Im}\, t(\mathbf{q}')$ ($t(\mathbf{q}')$ is transmission coefficient for wave vector $\mathbf{q}'$) for different incident wave vector $\mathbf{q} = S, M$, and K. For each incident wave vector $\mathbf{q}$, we checked nine transmission wave vectors $\mathbf{q}' = \mathbf{q} + \mathbf{G}'$, with $G' \in 0 \pm G_1^M, \pm G_2^M, \pm G_1^M \pm G_2^M$, labeled as $1 \sim 9(\mathbf{q}'_1 \sim \mathbf{q}'_9)$ in the illustration. The relative values of $-\mathrm{Im}\, t(\mathbf{q}')$ for undamped plasmon modes (corresponding to the peak value of $-\mathrm{Im}\, t(\mathbf{q}')$ at $15 \sim 20$ meV) are represented by the size of red circles.

system. This is equivalent to the result of the linear response theory, $\det \varepsilon(\mathbf{q}, \omega) = 0$, due to the relation $\det \mathcal{F}^{-1} = \det[\mathbf{V}(\mathbf{I} - \chi \mathbf{V})\mathbf{V}^{-1}] = \det(\mathbf{I} - \mathbf{V}\chi) = \det \varepsilon$. Moreover, the plasmon modes correspond to the peak values of transmission coefficients. For each $\mathbf{q}$, there are two peak values for transmission coefficients, corresponding to the damped plasmon mode and undamped plasmon mode, respectively, as depicted in Figs. 3(b)–3(d). The relative values of transmission coefficients for undamped plasmon modes are also represented by the size of circles. For $\mathbf{q} = S$, the plasmon modes are mainly contributed by the $\mathbf{q}' = \mathbf{q}$ and $\mathbf{q}' = \mathbf{q} - \mathbf{G}_2^M$ components labeled by 1 and 5, whereas the amplitudes of other wave-vector components are relatively small.

In Fig. 4(a), we have depicted the $\mathrm{Re}[\mathbf{E}_t(z)]$ distribution of the undamped plasmon modes for $\mathbf{q} = S$ within the xy plane. The electric field presents a 2D distribution within the xy plane, with significant periodic variations along the x direction and comparatively weaker periodic variations along the y direction. The periodicity of $\mathrm{Re}[\mathbf{E}_t(z)]$ in the x direction ($\lambda_1 = 30$ nm) aligns with the wave vector $\mathbf{q} = S$, indicating the contribution of $\mathbf{q}' = \mathbf{q}$ and $\mathbf{q}' = \mathbf{q} \pm \mathbf{G}_2^M$. In the y direction, the periodicity of $\mathrm{Re}[\mathbf{E}_t(z)]$ is $\lambda_2 = 8.5$ nm, suggesting the contributions of wave vectors $\mathbf{q}' = \mathbf{q} \pm \mathbf{G}_1^M$ and $\mathbf{q}' = \mathbf{q} \pm \mathbf{G}_1^M \pm \mathbf{G}_2^M$. Similar results can also be observed for a larger $\mathbf{q}$ which extends to the boundary of BZ, $\mathbf{q} = M$, as shown in Fig. 4(b), except that the $\mathrm{Re}[\mathbf{E}_t(z)]$ distribution

becomes more localized because of the larger $\mathbf{q}$ value. For the incident wave vector of $\mathbf{q} = K$, the undamped plasmon modes are contributed equally by $\mathbf{q}' = \mathbf{q}$, $\mathbf{q}' = \mathbf{q} + \mathbf{G}_1^M$, and $\mathbf{q}' = \mathbf{q} - \mathbf{G}_2^M$, as depicted in Fig. 3(d). The superposition of these electromagnetic waves results in a distinct 2D periodic distribution of $\mathrm{Re}[\mathbf{E}_t(z)]$ within the xy plane, as shown in Fig. 4(c). The emergence of these 2D $\mathrm{Re}[\mathbf{E}_t(z)]$ distribution patterns is closely correlated to the nonzero off-diagonal $\chi_{G,G'}$ components arising from inhomogeneous response which lifts the conservation of in-plane wave vectors. In the context of homogeneous response, the $\mathrm{Re}[\mathbf{E}_t(z)]$ would exhibit 1D distribution patterns along the direction parallel to $\mathbf{q}$, due to the conservation of in-plane wave vectors. Therefore, inhomogeneous optical response can be directly reflected by the distribution of plasmon electric fields in the real space, which can be detected by the near-field optical experiments. Moreover, the breakage of in-plane wave-vector conservation in inhomogeneous response can lead to anomalous optical scenarios, paving the way for innovative concept in the design of optoelectronic devices.

D. Nonlocal response in TTG

Finally, we demonstrate the nonlocal nature of the optical response in TTG. If this response is local, we would expect $\sigma_{\alpha\beta}(\mathbf{q}, \mathbf{q}', \omega) = \sigma_{\alpha\beta}(\mathbf{q} - \mathbf{q}', 0, \omega)$, resulting in

$\sigma_{\alpha\beta}(\mathbf{q}, \omega) = \sigma_{\alpha\beta}(\mathbf{q} \to 0, \omega)$ for $\mathbf{q} = \mathbf{q}'$. Conversely, a nonlocal response would yield $\sigma_{\alpha\beta}(\mathbf{q}, \omega) \neq \sigma_{\alpha\beta}(\mathbf{q} \to 0, \omega)$.

We calculated $\sigma_{xx}(\mathbf{q} \to 0, \omega)$ of TTG using the expression [48]

$$\sigma_{xx}(\mathbf{q} \to 0, \omega) = -\frac{ie^2 g\hbar}{(2\pi)^2} \sum_{\xi,m,n} \int \frac{(f_{\xi,\mathbf{k},m} - f_{\xi,\mathbf{k},n}) F_{m,n}^{\xi}(\mathbf{k})}{\left(E_{\mathbf{k},m}^{\xi} - E_{\mathbf{k},n}^{\xi}\right) \times \left(E_{\mathbf{k},m}^{\xi} - E_{\mathbf{k},n}^{\xi} + \hbar\omega + i\eta\right)} d^2\mathbf{k}, \tag{23}$$

with $F_{m,n}^{\xi}(\mathbf{k}) = \langle \xi, \mathbf{k}, m | v_x | \xi, \mathbf{k}, n \rangle \langle \xi, \mathbf{k}, n | v_x | \xi, \mathbf{k}, m \rangle$. In this expression, ξ is the valley index of TTG. For valley ξ, the energy eigenstates for wave vector $\mathbf{k}$ are defined as $|\xi, \mathbf{k}, m\rangle$ with the eigenenergy of $E_{\mathbf{k},m}^{\xi}$, where m is the band index, and $g = 2$ is the spin degeneracy. The conductivity $\sigma_{xx}(\mathbf{q}, \omega)$ at finite wave vector $\mathbf{q}$ can be determined by using the expression $\sigma_{xx}(\mathbf{q}, \omega) = i\chi(\mathbf{q}, \omega)e^2\omega/q^2$ for $\mathbf{q}$ along the x direction, according to Eq. (11).

Our calculations unveiled a significant distinction between $\sigma_{xx}(\mathbf{q} = S, \omega)$ and $\sigma_{xx}(\mathbf{q} \to 0, \omega)$ in TTG, as illustrated in Fig. 4(d), verifying the nonlocal features of the optical response in TTG. For frequency higher than 15 meV, $\text{Re}[\sigma_{xx}(\mathbf{q} \to 0, \omega)]$ converges to the value of graphene, $e^2/4\hbar$ [48], which can be attributed to the electron transitions in the DCs of TTG. In contrast, $\text{Re}[\sigma_{xx}(\mathbf{q} = S, \omega)]$ decreases to zero for $\hbar\omega > 10$ meV, due to the considerably larger wave vector

$\mathbf{q}$ compared to the momentum transfers for electron transitions in the DCs at these frequencies. The nonlocal response in TTG arises mainly from electronic transitions between moiré flatbands where quantum effects are pronounced, offering a promising avenue for investigating moiré flatbands using optical technique.

IV. CONCLUSIONS

In summary, we explored the plasmonic properties of magic-angle TTG and demonstrated their inhomogeneous and nonlocal response features, using the linear response theory and a continuous model Hamiltonian. Our findings unveil that the interlayer coupling, characterized by the moiré potential, plays a pivotal role in inducing significant inhomogeneous and nonlocal optical responses in the TTG. The inhomogeneity in optical responses can be manifest in the nonzero values of the off-diagonal elements of the extended dielectric matrix $\varepsilon_{\mathbf{G},\mathbf{G}'}$, which lead to exceptional characteristics of the plasmons arising from the electron transitions between the moiré flat bands of the TTG. Specifically, the inhomogeneity of optical responses leads to the violation of momentum conservation. Consequently, the conservation of in-plane wave vectors of electromagnetic waves is no longer applicable in the TTG. This phenomenon can also be visually represented through the special 2D distribution patterns of the plasmonic electric field, which offers an additional means for detecting the electronic states of materials by near-field optical experiments. Moreover, the intriguing optical scenarios that arise from these inhomogeneous and nonlocal responses present an alternative approach for advancing the development of optoelectronic devices.

FIG. 4. (a)–(c) The distribution of $\text{Re}\mathbf{E}_t(z)$ for the undamped plasmon modes in the xy and xz plane, for incident wave vectors $\mathbf{q} = S, M$, and K, respectively. The black arrows indicate the periodic of plasmon electric fields, and the green parallelograms denote the supercell of moiré lattice. (d) The real and imaginary parts of optical conductivities $\sigma_{xx}(\mathbf{q} = S, \omega)$ and $\sigma_{xx}(\mathbf{q} \to 0, \omega)$, as the functions of energy $\hbar\omega$.

APPENDIX A: RELATION BETWEEN DIELECTRIC FUNCTION MATRIX AND LATTICE SYMMETRY

Here, we demonstrate that the C_{2z} symmetry in the continuum model of TTG leads to $\chi(\mathbf{q}, \omega) = \chi(-\mathbf{q}, \omega)$ for arbitrary wave vector $\mathbf{q}$, and $\chi_{0,-\mathbf{G}_2^M}(\mathbf{q}, \omega) = \chi_{-\mathbf{G}_2^M,0}(\mathbf{q}, \omega)$ for $\mathbf{q} = M$.

Energy eigenstates $|\mathbf{k}, m\rangle$ and $|\mathbf{k} + \mathbf{q}, n\rangle$ with energies $E_{\mathbf{k},m}$ and $E_{\mathbf{k}+\mathbf{q},n}$ can be expanded as $|\mathbf{k}, m\rangle = S^{-1/2} \sum_{\mathbf{G}} c_{\mathbf{k},m}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}$ and $|\mathbf{k} + \mathbf{q}, n\rangle = S^{-1/2} \sum_{\mathbf{G}'} c_{\mathbf{k}+\mathbf{q},n}(\mathbf{G}') e^{i(\mathbf{k}+\mathbf{q}+\mathbf{G}')\cdot\mathbf{r}}$. The C_{2z} symmetry operator on a 2D system (in the xy plane) changes the wave vector $\mathbf{k}$ to $-\mathbf{k}$. Consequently, the energy eigenstates $|-\mathbf{k}, m\rangle$ and $|-\mathbf{k} - \mathbf{q}, n\rangle$ can
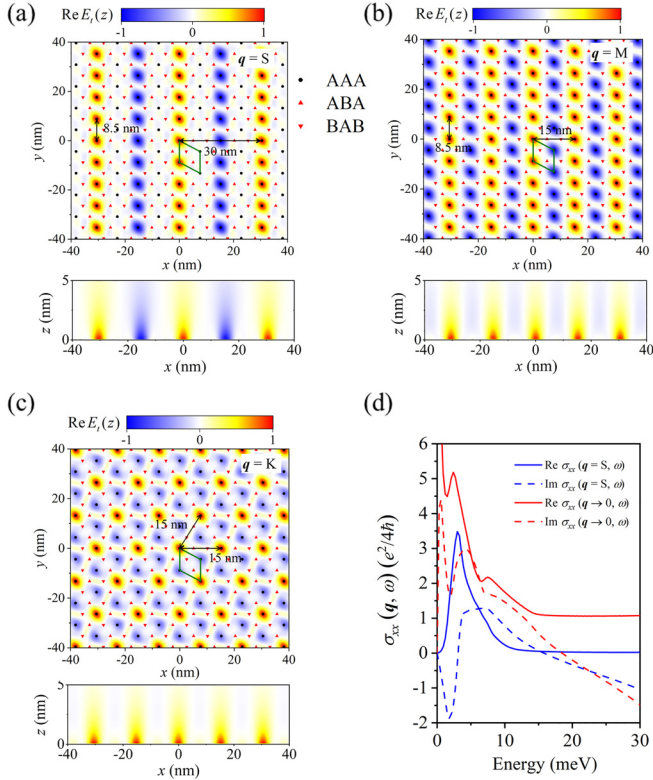
be expressed as $|-\mathbf{k}, m\rangle = S^{-1/2} \sum_{G} c_{\mathbf{k},m}(G) e^{-i(\mathbf{k}+G)\cdot\mathbf{r}}$ and $|-\mathbf{k}-\mathbf{q}, n\rangle = S^{-1/2} \sum_{G'} c_{\mathbf{k}+\mathbf{q},n}(G') e^{-i(\mathbf{k}+\mathbf{q}+G')\cdot\mathbf{r}}$.

We can arrive at $F_{k,m;k+q,n}(\mathbf{q}, \mathbf{q}) = F_{-k,m;-k-q,n}$ $(-\mathbf{q}, -\mathbf{q}) = |\sum_{\mathbf{G}} c^*_{\mathbf{k},m}(\mathbf{G}) c_{\mathbf{k}+\mathbf{q},n}(\mathbf{G})|^2$, which leads to $\chi(\mathbf{q}, \omega) = \chi(-\mathbf{q}, \omega)$ according Eq. (10).

Also, we can get the following expression:

$$F_{\mathbf{k},m;\mathbf{k}+\mathbf{q},n}\big(\mathbf{q}, \mathbf{q} - \mathbf{G}_2^M\big) = F_{-\mathbf{k},m;-\mathbf{k}-\mathbf{q},n}\big(-\mathbf{q}, -\mathbf{q} + \mathbf{G}_2^M\big)$$
$$= \left[\sum_{\mathbf{G}} c^*_{\mathbf{k},m}(\mathbf{G}) c_{\mathbf{k}+\mathbf{q},n}(\mathbf{G}) \right]$$
$$\times \left[\sum_{\mathbf{G}} c^*_{\mathbf{k}+\mathbf{q},n}\big(\mathbf{G} - \mathbf{G}_2^M\big) c_{\mathbf{k},m}(\mathbf{G}) \right]$$

Consequently, this expression results in $\chi(\mathbf{q}, \mathbf{q} - \mathbf{G}_2^M, \omega) = \chi(-\mathbf{q}, -\mathbf{q} + \mathbf{G}_2^M, \omega)$. For $\mathbf{q} = M$, we have $\mathbf{q} - \mathbf{G}_2^M = -\mathbf{q}$, and therefore, we have $\chi(\mathbf{q}, \mathbf{q} - \mathbf{G}_2^M, \omega) = \chi(\mathbf{q} - \mathbf{G}_2^M, \mathbf{q}, \omega)$.

For $\mathbf{q} = M$, and $\mathbf{G}, \mathbf{G}' \in \{0, -\mathbf{G}_2^M\}$, the dielectric function matrix elements in Eq. (18) are given as $\varepsilon_{11} = 1 - v(\mathbf{q})\chi(\mathbf{q}, \mathbf{q}, \omega)$, $\varepsilon_{22} = 1 - v(-\mathbf{q})\chi(-\mathbf{q}, -\mathbf{q}, \omega)$, $\varepsilon_{12} = -v(\mathbf{q})\chi(\mathbf{q}, \mathbf{q} - \mathbf{G}_2^M, \omega)$, and $\varepsilon_{21} = -v(-\mathbf{q})\chi(\mathbf{q} - \mathbf{G}_2^M, \mathbf{q}, \omega)$. Due to $v(\mathbf{q}) = v(-\mathbf{q})$, we can obtain that $\varepsilon_{11} = \varepsilon_{22}$ and $\varepsilon_{12} = \varepsilon_{21}$.

APPENDIX B: DERIVATION OF EQ. (22)

Maxwell's equation $\nabla \cdot \mathbf{E} = 0$ gives

$$iE_{i,\mathbf{q}+\mathbf{G}} - E_{i,z} = 0, \quad iE_{r,\mathbf{q}+\mathbf{G}'} + E_{r,z} = 0, \quad iE_{t,\mathbf{q}+\mathbf{G}'} - E_{t,z} = 0, \tag{B1}$$

Combining the charge-continuity equation and Eqs. (2) and (11), we can express the charge density as

$$\rho(\mathbf{q}+\mathbf{G}, \omega) = ie^2 \sum_{\mathbf{G}'} \chi(\mathbf{q}+\mathbf{G}, \mathbf{q}+\mathbf{G}', \omega) \frac{E_{t,\mathbf{q}+\mathbf{G}'}}{|\mathbf{q}+\mathbf{G}'|}. \tag{B2}$$

Using the boundary conditions $\mathbf{e}_z \times (\mathbf{E}_t - \mathbf{E}_i - \mathbf{E}_r)_{z=0} = 0$ and $\mathbf{e}_z \cdot (\mathbf{E}_t - \mathbf{E}_i - \mathbf{E}_r)_{z=0} = \rho/(\varepsilon_0\varepsilon)$, we get

$$E_{i,\mathbf{q}+\mathbf{G}} + E_{r,\mathbf{q}+\mathbf{G}} = E_{t,\mathbf{q}+\mathbf{G}}, \tag{B3}$$

$$i(E_{t,\mathbf{q}+\mathbf{G}} - E_{i,\mathbf{q}+\mathbf{G}} + E_{r,\mathbf{q}+\mathbf{G}}) = \frac{1}{\varepsilon_0\varepsilon}\rho(\mathbf{q}+\mathbf{G}, \omega). \tag{B4}$$

Combining Eqs. (B2)–(B4), we can arrive at

$$\sum_{\mathbf{G}'} [\delta_{\mathbf{G},\mathbf{G}'} - v(\mathbf{q}+\mathbf{G}')\chi_{\mathbf{G},\mathbf{G}'}(\mathbf{q}, \omega)]E_{t,\mathbf{q}+\mathbf{G}'} = E_{i,\mathbf{q}+\mathbf{G}}, \tag{B5}$$

where we define $v(\mathbf{q}+\mathbf{G}') = \frac{e^2}{2\varepsilon_0\varepsilon|\mathbf{q}+\mathbf{G}'|}$ and $\chi_{\mathbf{G},\mathbf{G}'}(\mathbf{q}, \omega) = \chi(\mathbf{q}+\mathbf{G}, \mathbf{q}+\mathbf{G}', \omega)$. Equation (B5) can also be written as a simple equation, Eq. (22).

[1] F. Bonaccorso, Z. Sun, T. Hasan, and A. C. Ferrari, Graphene photonics and optoelectronics, Nat. Photon. **4**, 611 (2010).

[2] A. B. Kuzmenko, E. van Heumen, F. Carbone, and D. van der Marel, Universal optical conductance of graphite, Phys. Rev. Lett. **100**, 117401 (2008).

[3] K. F. Mak, M. Y. Sfeir, Y. Wu, C. H. Lui, J. A. Misewich, and T. F. Heinz, Measurement of the optical conductivity of graphene, Phys. Rev. Lett. **101**, 196405 (2008).

[4] P. Matyba, H. Yamaguchi, G. Eda, M. Chhowalla, L. Edman, and N. D. Robinson, Graphene and mobile ions: The key to all-plastic, solution-processed light-emitting devices, ACS Nano **4**, 637 (2010).

[5] Z. Sun, T. Hasan, F. Torrisi, D. Popa, G. Privitera, F. Wang, F. Bonaccorso, D. M. Basko, and A. C. Ferrari, Graphene mode-locked ultrafast laser, ACS Nano **4**, 803 (2010).

[6] F. Xia, T. Mueller, Y.-m. Lin, A. Valdes-Garcia, and P. Avouris, Ultrafast graphene photodetector, Nat. Nanotechnol. **4**, 839 (2009).

[7] C. Ding and M. Zhao, Chiral response in two-dimensional bilayers with time-reversal symmetry: A universal criterion, Phys. Rev. B **108**, 125415 (2023).

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

[9] N. A. Mortensen, S. Raza, M. Wubs, T. Søndergaard, and S. I. Bozhevolnyi, A generalized non-local optical response theory for plasmonic nanostructures, Nat. Commun. **5**, 3809 (2014).

[10] M. B. Lundeberg, Y. Gao, R. Asgari, C. Tan, B. Van Duppen, M. Autore, P. Alonso-González, A. Woessner, K. Watanabe, T. Taniguchi, R. Hillenbrand, J. Hone, M. Polini, and F. H. L. Koppens, Tuning quantum nonlocal effects in graphene plasmonics, Science **357**, 187 (2017).

[11] W. L. Barnes, A. Dereux, and T. W. Ebbesen, Surface plasmon subwavelength optics, Nature (London) **424**, 824 (2003).

[12] S. Vedantam, H. Lee, J. Tang, J. Conway, M. Staffaroni, and E. Yablonovitch, A plasmonic dimple lens for nanoscale focusing of light, Nano Lett. **9**, 3447 (2009).

[13] A. Karalis, E. Lidorikis, M. Ibanescu, J. D. Joannopoulos, and M. Soljačić, Surface-plasmon-assisted guiding of broadband slow and subwavelength light in air, Phys. Rev. Lett. **95**, 063901 (2005).

[14] S. V. Boriskina, H. Ghasemi, and G. Chen, Plasmonic materials for energy: From physics to applications, Mater. Today **16**, 375 (2013).

[15] K. M. Mayer and J. H. Hafner, Localized surface plasmon resonance sensors, Chem. Rev. **111**, 3828 (2011).

[16] H. A. Atwater and A. Polman, Plasmonics for improved photovoltaic devices, Nat. Mater. **9**, 205 (2010).

[17] C. Clavero, Plasmon-induced hot-electron generation at nanoparticle/metal-oxide interfaces for photovoltaic and photocatalytic devices, Nat. Photon. **8**, 95 (2014).

[18] V. J. Sorger, R. F. Oulton, R.-M. Ma, and X. Zhang, Toward integrated plasmonic circuits, MRS Bull. **37**, 728 (2012).

[19] J. Chen, M. Badioli, P. Alonso-González, S. Thongrattanasiri, F. Huth, J. Osmond, M. Spasenović, A. Centeno, A. Pesquera, P. Godignon, A. Zurutuza Elorza, N. Camara, F. J. G. de Abajo, R. Hillenbrand, and F. H. L. Koppens, Optical nano-imaging of gate-tunable graphene plasmons, Nature (London) **487**, 77 (2012).

[20] F. H. L. Koppens, D. E. Chang, and F. J. García de Abajo, Graphene plasmonics: A platform for strong light–matter interactions, Nano Lett. **11**, 3370 (2011).

[21] L. Brey, T. Stauber, L. Martín-Moreno, and G. Gómez-Santos, Nonlocal quantum effects in plasmons of graphene superlattices, Phys. Rev. Lett. **124**, 257401 (2020).

[22] R. Bistritzer and A. H. MacDonald, Moiré bands in twisted double-layer graphene, Proc. Natl. Acad. Sci. USA **108**, 12233 (2011).

[23] M. Koshino, N. F. Q. Yuan, T. Koretsune, M. Ochi, K. Kuroki, and L. Fu, Maximally localized Wannier orbitals and the extended Hubbard model for twisted bilayer graphene, Phys. Rev. X **8**, 031087 (2018).

[24] Y. Cao, V. Fatemi, A. Demir, S. Fang, S. L. Tomarken, J. Y. Luo, J. D. Sanchez-Yamagishi, K. Watanabe, T. Taniguchi, E. Kaxiras, R. C. Ashoori, and P. Jarillo-Herrero, Correlated insulator behaviour at half-filling in magic-angle graphene superlattices, Nature (London) **556**, 80 (2018).

[25] Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero, Unconventional superconductivity in magic-angle graphene superlattices, Nature (London) **556**, 43 (2018).

[26] A. Kerelsky, L. J. McGilly, D. M. Kennes, L. Xian, M. Yankowitz, S. Chen, K. Watanabe, T. Taniguchi, J. Hone, C. Dean, A. Rubio, and A. N. Pasupathy, Maximized electron interactions at the magic angle in twisted bilayer graphene, Nature (London) **572**, 95 (2019).

[27] Y. Choi, J. Kemmer, Y. Peng, A. Thomson, H. Arora, R. Polski, Y. Zhang, H. Ren, J. Alicea, G. Refael, F. von Oppen, K. Watanabe, T. Taniguchi, and S. Nadj-Perge, Electronic correlations in twisted bilayer graphene near the magic angle, Nat. Phys. **15**, 1174 (2019).

[28] X. Lu, P. Stepanov, W. Yang, M. Xie, M. A. Aamir, I. Das, C. Urgell, K. Watanabe, T. Taniguchi, G. Zhang, A. Bachtold, A. H. MacDonald, and D. K. Efetov, Superconductors, orbital magnets and correlated states in magic-angle bilayer graphene, Nature (London) **574**, 653 (2019).

[29] Y. Xie, B. Lian, B. Jäck, X. Liu, C.-L. Chiu, K. Watanabe, T. Taniguchi, B. A. Bernevig, and A. Yazdani, Spectroscopic signatures of many-body correlations in magic-angle twisted bilayer graphene, Nature (London) **572**, 101 (2019).

[30] M. Yankowitz, S. Chen, H. Polshyn, Y. Zhang, K. Watanabe, T. Taniguchi, D. Graf, A. F. Young, and C. R. Dean, Tuning superconductivity in twisted bilayer graphene, Science **363**, 1059 (2019).

[31] C. Lewandowski and L. Levitov, Intrinsically undamped plasmon modes in narrow electron bands, Proc. Natl. Acad. Sci. USA **116**, 20869 (2019).

[32] T. Stauber and H. Kohler, Quasi-flat plasmonic bands in twisted bilayer graphene, Nano Lett. **16**, 6844 (2016).

[33] P. Novelli, I. Torre, F. H. L. Koppens, F. Taddei, and M. Polini, Optical and plasmonic properties of twisted bilayer graphene: Impact of interlayer tunneling asymmetry and ground-state charge inhomogeneity, Phys. Rev. B **102**, 125403 (2020).

[34] X. Kuang, Z. Zhan, and S. Yuan, Collective excitations and flat-band plasmon in twisted bilayer graphene near the magic angle, Phys. Rev. B **103**, 115431 (2021).

[35] C. Ding, X. Zhang, H. Gao, X. Ma, Y. Li, and M. Zhao, Role of electron-electron interaction in the plasmon modes of twisted bilayer graphene, Phys. Rev. B **106**, 155402 (2022).

[36] L. Brey, T. Stauber, T. Slipchenko, and L. Martín-Moreno, Plasmonic Dirac cone in twisted bilayer graphene, Phys. Rev. Lett. **125**, 256804 (2020).

[37] N. C. H. Hesp, I. Torre, D. Rodan-Legrain, P. Novelli, Y. Cao, S. Carr, S. Fang, P. Stepanov, D. Barcons-Ruiz, H. Herzig Sheinfux, K. Watanabe, T. Taniguchi, D. K. Efetov, E. Kaxiras, P. Jarillo-Herrero, M. Polini, and F. H. L. Koppens, Observation of interband collective excitations in twisted bilayer graphene, Nat. Phys. **17**, 1162 (2021).

[38] F. Hu, S. R. Das, Y. Luan, T. F. Chung, Y. P. Chen, and Z. Fei, Real-space imaging of the tailored plasmons in twisted bilayer graphene, Phys. Rev. Lett. **119**, 247402 (2017).

[39] C. Ding, H. Gao, L. Sun, X. Ma, and M. Zhao, Corrugation effect, Dirac cone splitting, and plasmon properties of biased twisted bilayer graphene, Phys. Rev. B **104**, 155427 (2021).

[40] M. Papaj and C. Lewandowski, Probing correlated states with plasmons, Sci. Adv. **9**, eadg3262 (2023).

[41] S. Carr, C. Li, Z. Zhu, E. Kaxiras, S. Sachdev, and A. Kruchkov, Ultraheavy and ultrarelativistic Dirac quasiparticles in sandwiched graphenes, Nano Lett. **20**, 3030 (2020).

[42] J. M. Park, Y. Cao, K. Watanabe, T. Taniguchi, and P. Jarillo-Herrero, Tunable strongly coupled superconductivity in magic-angle twisted trilayer graphene, Nature (London) **590**, 249 (2021).

[43] E. Khalaf, A. J. Kruchkov, G. Tarnopolsky, and A. Vishwanath, Magic angle hierarchy in twisted graphene multilayers, Phys. Rev. B **100**, 085109 (2019).

[44] Z. Wu, X. Kuang, Z. Zhan, and S. Yuan, Magic angle and plasmon mode engineering in twisted trilayer graphene with pressure, Phys. Rev. B **104**, 205104 (2021).

[45] T. Cea and F. Guinea, Coulomb interaction, phonons, and superconductivity in twisted bilayer graphene, Proc. Natl. Acad. Sci. USA **118**, e2107874118 (2021).

[46] P. A. D. Gonçalves and N. M. R. Peres, *An Introduction to Graphene Plasmonics* (World Scientific, 2016).

[47] A. Otto, Excitation of nonradiative surface plasma waves in silver by the method of frustrated total reflection, Z. Physik **216**, 398 (1968).

[48] S. A. Mikhailov and K. Ziegler, New electromagnetic mode in graphene, Phys. Rev. Lett. **99**, 016803 (2007).