Electronic-Based Model of the Optical Nonlinearity of Low-Electron-Density Drude Materials

Ieng-Wai Un[*,§] Subhajit Sarkar[†,§] and Yonatan Sivan[‡]

School of Electrical and Computer Engineering, Ben-Gurion University of the Negev and the Ilse Katz Center for Nanoscale Science and Technology, Ben-Gurion University of the Negev, Beer Sheva, Israel

Low-electron-density Drude (LEDD) materials such as indium tin oxide (ITO) are receiving considerable attention for their combination of CMOS compatibility, unique epsilon-near-zero (ENZ) behavior, and giant ultrafast nonlinear thermo-optic response. However, current understanding of the electronic and optical response of LEDD materials is limited due to the simplistic modeling that extends only the known models of noble metals without considering the interplay among the lower electron density, relatively high Debye energy, and the nonparabolic band structure. We bridge this knowledge gap and provide a complete understanding of the nonlinear electronic-thermal-optical response of LEDD materials. In particular, we rely on state-of-the-art electron dynamics modeling, as well as a time-dependent permittivity model for LEDD materials under optical pumping within the adiabatic approximation. We find the electron temperature may reach values much higher than realized before, even exceeding the Fermi temperature, in which case the effective chemical potential dramatically decreases and even becomes negative, thus, transient giving the material some characteristics of a semiconductor. We further show that the nonlinear optical response of LEDD materials originating from the changes to the real part of the permittivity is associated with changes of the population. This resolves the argument about the rise time of the permittivity, showing that it is instantaneous. In this vein, we show that referring to the LEDD permittivity as having a Kerr or "saturable" nonlinearity is unsuitable since its permittivity dynamics is absorptive rather than nonresonant and does not originate from population inversion. Finally, we analyze the probe-pulse dynamics and unlike previous work, we obtain a quantitative agreement with the results of recent experiments.

I. INTRODUCTION

The technologically relevant transparent conducting oxide ITO has recently been shown to possess an exceptionally strong ultrafast response to illumination, making it a promising candidate for nonlinear optics applications. Its strong response was initially associated with the unique epsilon-near-zero (ENZ) point it has in the near infrared [1] (see also Fig. 1), which is characteristic of low-electron-density Drude (LEDD) materials.

The nonlinear subpicosecond optical response of LEDD materials was described using a variety of approaches, starting from a phenomenological temporally delayed response or as having a time-dependent effective mass (and hence, plasma frequency), see, e.g., Refs. [2–9]. The latter is a simple description of the transient occupation of high-energy electron states for which the effective mass is higher than for low-energy states.

More recent models employed a relaxation time approximation- (RTA) based extended two-temperature model (ETTM) [4,9–11]. Notably, while such models offered qualitative agreement with measurements, reaching also a quantitative match required various ad hoc corrections see, e.g., Ref. [9]. This could have originated from the use of equations derived for parabolic bands, because the significance of momentum conservation to the e-ph interactions was not yet understood [12], because the e-e interaction were accounted for in a simplistic manner [13], because the effect of the phonon temperature on the ITO permittivity was not always accounted for, or because the analysis of the nonlinear response was specific for every experimental configuration. Thus, for example, the computed ETTM parameters in Ref. [9] (including the strength of the e-ph interaction and the T_e dependence of the damping factor) had to be manually changed to fit the experimental data, without a firm theoretical justification.

Notably, there are two additional limitations of the state-of-the-art nonlinear optical theory of LEDD materials. First, a significant aspect of the dynamics that is not yet well understood is the response to high illumination intensities. In Refs. [4,8,14], this response was referred

---

*iengwai@post.bgu.ac.il

†subhajit@post.bgu.ac.il

‡sivanyon@bgu.ac.il

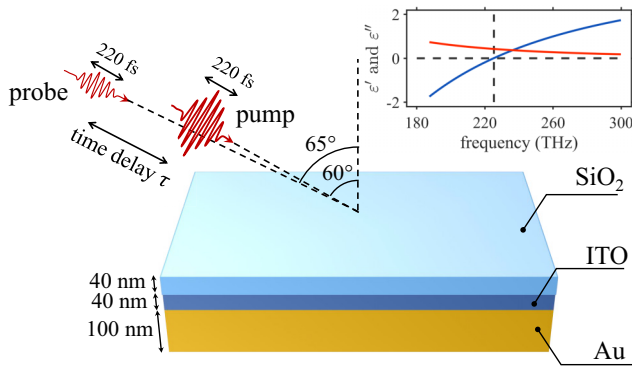§These two authors have contributed equally.

FIG. 1. Schematic diagram of the setup considered here and in experiment [8]. The sample consists of a 40-nm thin film of ITO deposited on glass and covered by a 100-nm gold film. A pump and a probe pulse are obliquely incident on the sample at angles of 60° and 65°, respectively. The probe pulse arrives at the sample with a time delay τ after the pump pulse. The inset shows the real (blue solid line) and imaginary (red solid line) parts of the ITO permittivity at 300 K (i.e., plug in the Fermi-Dirac distribution at 300 K to Eq. (9)). The black dashed lines indicate the ENZ point at approximately 225 THz.

to as saturable; the saturation intensity was claimed to be approximately 100 GW/cm² [8], but this value was not connected to a population inversion (as in atomic media) nor to a balance of excitation and relaxation mechanisms, as done for Drude metals [15]. Indeed, usually, a nonlinear saturable response is associated with the gradual depletion of the (electronic) ground state of the material, an effect that leads to a *decrease* of the imaginary part of its permittivity (saturable absorption) [16], and even to a change of its sign (upon population inversion) [17,18]. This behavior is in contradiction to the experimental results that show that the imaginary part of the ITO permittivity *grows* upon illumination, see, e.g., Refs. [9,14]. Thus, to understand whether the response of LEDD materials is saturable or not requires simultaneous monitoring of the population as well as of the imaginary part of the permittivity via electronic simulations.

Second, the strong changes of the permittivity necessarily lead to changes of the local field, which then affects the distribution and permittivity and vice versa. To date, the coupling between these properties was not treated self-consistently.

A step towards resolution of the above questions was made in Ref. [12], where the electron distribution dynamics was studied by solving the Boltzmann equation and extracting from it the (effective) electron and phonon temperatures, as well as the underlying thermal properties such as the heat capacity and e-ph coupling. This work revealed the significance of momentum conservation in e-ph interactions, the faster e-e collision rates, and large quantitative differences in the values of various parameters compared

to noble metals. Notably, however, the simulations in Ref. [12] were limited to modestly high intensities.

In this work, we go beyond Ref. [12] and provide a detailed electronic model of the optical properties of LEDD materials and apply it to illumination levels reaching the reported damage threshold. In Sec. II, we recall briefly the electronic model developed in Ref. [12], and describe the time-dependent permittivity model used in this work. In particular, we model the permittivity by applying the Lindhard formula to the nonparabolic band characteristic of LEDD materials and describe a self-consistent approach for the calculation of the distribution, permittivity, and local field.

As a specific example, we then focus on a prototypical geometry of an ITO layer illuminated by an obliquely incident pump pulse. In Sec. III A, we discuss the resulting dynamics of the electron and phonon temperatures as a function of illumination intensity. We show that the electron temperature may reach values much higher than realized before, even exceeding the Fermi temperature, and the chemical potential may become negative, giving the material some transient characteristics of a semiconductor. We also show that the decay rate of the electron temperature becomes faster with the illumination intensity.

In Sec. III B, we study the permittivity dynamics in detail, including the various contributions to the change of its real and imaginary parts. We show that the dynamics is dominated by the change of the real part of the permittivity (causing a frequency shift of the ENZ point), which itself is dominated by changes to the electron distribution. The increase of the imaginary part of the permittivity is also large, but not nearly as much as that of the real part; its dynamics results from (sometimes opposite) contributions of various effects and both electron and phonon temperatures. The former insight explains the observation of the slower decay rate of the reflectivity with intensity reported in Ref. [8]. Of note, these results show that the ITO permittivity dynamics has an unusual absorptive yet instantaneous response, which is responsible for the turn on stage of the nonlinearity, whereas the turn off stage has a thermal nature, yet a relatively fast one (in agreement with Ref. [19]). Our results also show that the response has some characteristics of a high-quality metal (as discussed in Ref. [20] for pulsed illumination, and even in Refs. [21,22] for cw illumination), so that although the local field grows sublinearly with illumination intensity, the ITO is *not* a saturable absorber (for which the real part of the permittivity hardly changes and its imaginary part decreases with growing illumination intensity, but rather exhibits the opposite behavior). By comparing our simulation results with the two-temperature model (TTM), we show that the electron dynamics described by the TTM is an excellent approximation for calculating the nonlinear optical response of ITO when the pulse duration is longer than the relatively short thermalization time. We

also explain the physical origin of the ad hoc corrections needed previously, e.g., in Ref. [9]. Then, in Sec. III C, we discuss the energy partition and show that our prediction for the intensity at which the phonon temperature reaches the melting temperature matches well the experimentally reported damage threshold. In Sec. III D we study the probe-pulse dynamics. In particular, we show that the decay rate of the reflection of the probe pulse decreases with the pump-pulse intensity due to the frequency shift of the ENZ point. This result provides quantitative agreement with the experimental observation in Ref. [8]. We also study the nonlinear optical response to shorter pulses in Sec. III E and show that one needs to go beyond the TTM only when the pulse duration is shorter than the thermalization time, by accounting for the nonthermal part of the distribution. Finally, we provide a discussion of the results in Sec. IV and the conclusion in Sec. V.

II. FORMULATION

A. Model for the electron distribution dynamics and phonon temperature

We start by revisiting and extending the theoretical approach presented in Ref. [12]. In particular, we solve the Boltzmann equation (BE) for the electron dynamics, namely,

$$
\begin{aligned}
\frac{\partial f(\mathcal{E}, t)}{\partial t} &= \left(\frac{\partial f(\mathcal{E}, t; \varepsilon(t, \omega_{\text{pump}}), |\mathbf{E}(t, \omega_{\text{pump}})|^2)}{\partial t} \right)_{\text{exc}} \\
&+ \left(\frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{e\text{-ph collision}} \\
&+ \left(\frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{e\text{-}e \text{ collision}} + \left(\frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{e\text{-imp collision}}.
\end{aligned}
\tag{1}
$$

Here, $f(\mathcal{E}, t)$ is the electron distribution function at an energy $\mathcal{E}$ and time t, representing the population probability of electrons in a system characterized by a continuum of electron energy states within the conduction band. While the expressions for the electron-electron (e-e), electron-phonon (e-ph), and electron-charged-impurity (e-imp) interactions are the same as in Refs. [12,23–25], the excitation term is more complicated. In particular, the electron population evolves in time such that it is given by

$$
\begin{aligned}
\left(\frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{\text{exc}} &= B(t; \omega_{\text{pump}}) \Big[D_J(\mathcal{E} - \hbar\omega_{\text{pump}}, \mathcal{E}) \rho_e(\mathcal{E} - \hbar\omega_{\text{pump}}) f(\mathcal{E} - \hbar\omega_{\text{pump}}, t)) (1 - f(\mathcal{E}, t)) \\
&\quad - D_J(\mathcal{E}, \mathcal{E} + \hbar\omega_{\text{pump}}) \rho_e(\mathcal{E} + \hbar\omega_{\text{pump}}) f(\mathcal{E}, t)) (1 - f(\mathcal{E} + \hbar\omega_{\text{pump}}, t)) \Big],
\end{aligned}
\tag{2}
$$

where $D_J(\mathcal{E}_{\text{initial}}, \mathcal{E}_{\text{final}})$ is the squared magnitude of the transition matrix element for the electronic process $\mathcal{E}_{\text{initial}} \to \mathcal{E}_{\text{final}}$. The constant $B(t; \omega_{\text{pump}})$ is determined by ensuring that the increase rate of the energy density of the electron subsystem due to the excitation is equal to the absorbed power density p_{abs}, namely,

$$
\int \mathcal{E} \rho_e(\mathcal{E}) \left(\frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{\text{exc}} d\mathcal{E} = p_{\text{abs}}(t; \omega_{\text{pump}}).
\tag{3}
$$

Here, the absorbed power density is evaluated dynamically and self-consistently with $f(\mathcal{E}, t)$, $\varepsilon(t; \omega_{\text{pump}})$, and $\mathbf{E}(t)$ using Poynting's theorem (see details in Sec. II B).

The BE (2) is complemented by a coarse-grained equation for the phonon temperature, which follows from energy exchange between the electron and phonon subsystems (as done previously in Refs. [25,26]), see Appendix B; energy transfer to the environment occurs on the time scale of many picoseconds, thus, can be ignored in the context of the ultrafast dynamics.

These coupled equations are now used to determine the permittivity dynamics.

B. Self-consistent field calculations and model for a time-dependent permittivity

When the pump-pulse energy is high enough to induce a non-negligible dynamical change in the electron distribution, the electron system cannot be considered to be time translation invariant. In this case, one cannot apply the convolution theorem to the relation between the polarization vector and the electric field vector. Strictly speaking, this requires one to solve Maxwell's equations coupled with the Boltzmann equation using a time step much finer than the periodicity of the electromagnetic wave. Such a calculation can be very time consuming and computationally expensive [27]. However, as shown in Appendix A, this problem can be circumvented by applying the adiabatic approximation to calculate the time evolution of the electric field when the instantaneous intensity of the incoming

pulse is weak enough such that the change rate of the electron distribution is slower than the damping rate. This condition is satisfied when the local field is smaller than approximately 2.7×10^9 V/m. For this purpose, we further assume that the pump-pulse duration is much longer than the periodicity so that the electric field $\vec{\mathcal{E}}(t)$ [and the electric displacement $\vec{\mathcal{D}}(t)$] can be considered quasi-monochromatic and be written as a product of a slowly varying envelope [$\mathbf{E}(t)$ or $\mathbf{D}(t)$] and a rapidly varying phase factor, namely, $\vec{\mathcal{E}}(t) = \mathbf{E}(t)e^{-i\omega_{\text{pump}}t} + \text{c.c.}$ ($\vec{\mathcal{D}}(t) =$

$\mathbf{D}(t)e^{-i\omega_{\text{pump}}t} + \text{c.c.}$). We show in Appendix A that the electric displacement envelope is then related to the local field envelope by $\mathbf{D}(t) = \varepsilon_0 \varepsilon(t, \omega_{\text{pump}})\mathbf{E}(t)$, and that Ampère's law and the Maxwell-Faraday equation become

$$\begin{cases} \nabla \times \mathbf{H}(\mathbf{r}, t) = -i\omega_{\text{pump}}\varepsilon_0 \varepsilon(t; \omega_{\text{pump}})\mathbf{E}(\mathbf{r}, t) \\ \nabla \times \mathbf{E}(\mathbf{r}, t) = -i\omega_{\text{pump}}\mu_0 \mathbf{H}(\mathbf{r}, t) \end{cases}. \quad (4)$$

Here, $\varepsilon(t; \omega_{\text{pump}})$ is the time-dependent permittivity given by

$$\varepsilon(t; \omega_{\text{pump}}) = \varepsilon_\infty + \lim_{\mathbf{q} \to 0} \frac{2e^2}{\varepsilon_0 q^2} \int \frac{d^3k}{(2\pi)^3} \frac{f_{\mathbf{k}+\mathbf{q}}(t; \omega_{\text{pump}}) - f_{\mathbf{k}}(t; \omega_{\text{pump}})}{\mathcal{E}_{\mathbf{k}+\mathbf{q}} - \mathcal{E}_{\mathbf{k}} - \hbar\omega_{\text{pump}} - i\hbar(\eta_{\mathbf{k}+\mathbf{q}}(t; \omega_{\text{pump}}) + \eta_{\mathbf{k}}(t; \omega_{\text{pump}}))/2}, \quad (5)$$

where ε_∞ represents the contribution of interband transitions to the permittivity, $f_{\mathbf{k}}$ is the momentum (**k**)-dependent electron distribution in the conduction band, $\mathcal{E}_{\mathbf{k}}$ is the electron energy satisfying the $\mathcal{E}$-k relation [see Eq. (7) below], **q** is the wave vector of the applied field (taken to the zero limit because it is much smaller than the electron wave vector), and $\eta_{\mathbf{k}} = \tau_{e\text{-}e,\mathbf{k}}^{-1} + \tau_{e\text{-}ph,\mathbf{k}}^{-1} + \tau_{e\text{-}imp,\mathbf{k}}^{-1}$ is the damping rate following from Matthiessen's rule [28]. $\tau_{e\text{-}e,\mathbf{k}}^{-1}$, $\tau_{e\text{-}ph,\mathbf{k}}^{-1}$, and $\tau_{e\text{-}imp,\mathbf{k}}^{-1}$ are the collision rates associated with the e-e, e-ph, and e-imp interactions, respectively (see details in Ref. [12]). In addition, the contribution of the interband transition (ε_∞) is assumed to be independent of time, temperature, and electric field and not to contribute to the dispersion at frequencies below the interband threshold of approximately 3.2 eV.

The local field envelope [the solution of Eq. (4)] and the time-dependent permittivity (5) are next used to calculate the power absorbed density via Poynting's theorem

$$p_{\text{abs}}(t; \omega_{\text{pump}}) = \frac{\omega_{\text{pump}}}{2}\varepsilon_0 \varepsilon''(t; \omega_{\text{pump}})|\mathbf{E}(t; \omega_{\text{pump}})|^2. \quad (6)$$

Then, we substitute the absorbed power density back to Eqs. (2) and (3) to evaluate the electron-photon excitation term so that the electron dynamics and the pump-pulse dynamics are solved self-consistently. Consequently, we can define Eqs. (1)–(6) as an adiabatic nonthermal permittivity ($\mathcal{E}$) model (ANTH$\mathcal{E}$M).

It is then customary to rewrite the integral in energy space. For materials with a nonparabolic conduction band, e.g., in ITO, the energy-momentum ($\mathcal{E}$-k) relation is expressed by the Kane quasilinear dispersion [29],

$$\hbar^2 k^2 = 2m_e^* \mathcal{E}_{\mathbf{k}}(1 + C\mathcal{E}_{\mathbf{k}}), \quad (7)$$

such that the electron density of states (EDOS) is given by [12,29,30]

$$\rho_e(\mathcal{E}) = \frac{1 + 2C\mathcal{E}}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2}\sqrt{\mathcal{E}(1 + C\mathcal{E})}, \quad (8)$$

where $m_e^* = 0.3964m_e$ is the electron effective mass at the conduction-band minimum, and $C = 0.4191$ eV^{-1} [30] is the first-order nonparabolicity factor. By substituting the $\mathcal{E}$-k relation into Eq. (5), expressing the integrand as a power series of q, and converting the integral over **k** to an integral over $\mathcal{E}$ using the EDOS (8), we obtain

$$\varepsilon(t; \omega_{\text{pump}}) = \varepsilon_\infty - \frac{e^2}{\varepsilon_0 m_e^*}\int \frac{\rho_e(\mathcal{E})f(\mathcal{E}, t; \omega_{\text{pump}})}{(\omega_{\text{pump}} + i\eta(\mathcal{E}, t; \omega_{\text{pump}}))^2} \frac{(1 + 8C\mathcal{E}(1 + C\mathcal{E})/3)}{(1 + 2C\mathcal{E})^3}d\mathcal{E}. \quad (9)$$

In the context of ITO modeling, Eq. (9) represents a rigorous and unique extension of Lindhard's formula to time-varying systems having a nonparabolic conduction band. In particular, the permittivity, Eq. (9), includes the contribution from the nonthermal part of the electron distribution, instead of the thermal part only, as in all previous work. We show below that the deviation from equilibrium is negligible for pulse durations much longer than the e-e relaxation time $\tau_{e\text{-}e}$ such as those used in Refs. [2–5,7–9], but that it is non-negligible for pulse durations comparable or shorter than $\tau_{e\text{-}e}$. Moreover, the damping rate η in Eq. (9) is obtained self-consistently from the electron distribution via the e-e, e-ph, and e-imp scattering rates (see Ref. [12]). This is in contrast with the phenomenological methods addressing the permittivity damping term in previous studies [3,4,9,11,31]. Below, we show that this

self-consistent form successfully explains the experimental data [8,10,14] without introducing phenomenological adjustments.

If the damping term η is approximated to be energy-independent, then, Eq. (9) reproduces the Drude formula for the relative permittivity with an electron-distribution-dependent plasma frequency [3,4,9,11,31], namely,

$$\omega_p^2[f(\mathcal{E})] = \frac{e^2}{3\pi^2\varepsilon_0 m_e^*} \int \left[\frac{2m_e^*}{\hbar^2}\mathcal{E}(1+C\mathcal{E})\right]^{3/2}$$

$$(1+2C\mathcal{E})^{-1}\left(-\frac{\partial f}{\partial\mathcal{E}}\right)d\mathcal{E}. \qquad (10)$$

The extra factor for $C \neq 0$ in the integral of Eqs. (9) and (10) originates from the nonparabolic energy-momentum relation of ITO when expanding the denominator of Eq. (5) in a power series of q; it reflects the energy dependence of the electron effective mass due to the nonparabolicity.

Finally, if the electron distribution function can be approximated by a thermal (i.e., Fermi-Dirac) distribution, then, the (dynamics of the) permittivity can be described as a function of the electron and phonon temperatures only. Here, the electron and phonon temperatures can be obtained by solving the TTM (see Appendix B).

C. Heuristic explanation of temperature and permittivity dynamics

Before dwelling into the detailed description of the temperature and permittivity dynamics, it is useful to provide an intuitive and simple description of those as a function of the illumination intensity. In particular, the maximum of the electron and phonon temperatures and local fields can be qualitatively understood by the illumination-induced ENZ resonance shift.

The absorption of the pump-pulse energy results in an increase of the electron temperature [see Fig. 2(a) below] and thus an increase in the real part of the permittivity [see Fig. 4(a) below]. This causes the ENZ resonance to slip away from the incoming frequencies so that the local field and the absorptivity decrease (see Fig. 5); consequently, the total absorbed energy, the maximum electron and phonon temperature increases sublinearly (rather than linearly) with the illumination intensity. As shown below, this effect explains the high damage threshold of ITO. The changes of the imaginary part of the ITO permittivity are also large, and similar in nature to those occurring in noble metals (see, e.g. Refs. [21,22]), but are secondary compared to the large ENZ resonance shift.

III. RIGOROUS RESULTS

As a specific example, we consider a system similar to the one studied in Ref. [8], where a pump and a probe pulse are obliquely incident on a sample consisting of a
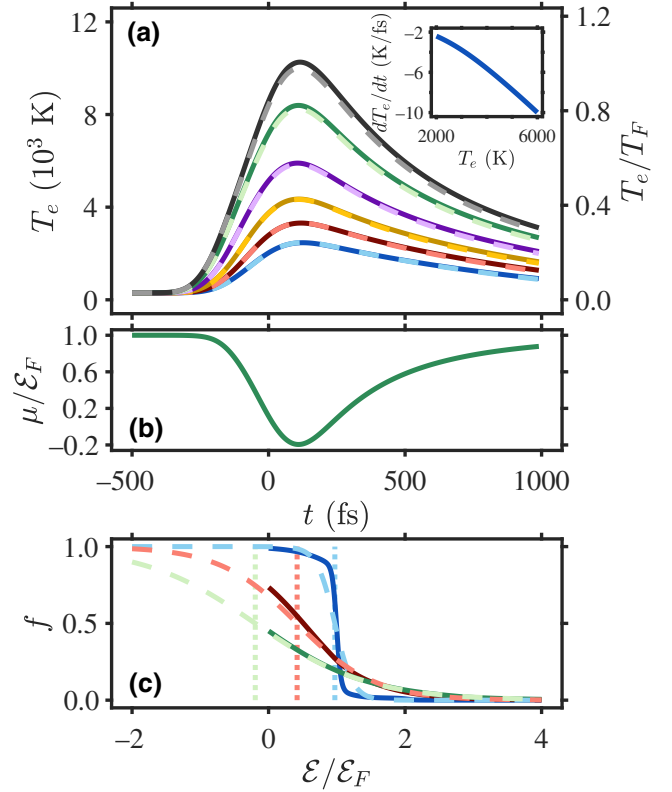

FIG. 2. (a) Electron temperature (normalized to the Fermi temperature $T_F = 10204$ K on the right axis) as a function of time for pulse intensity $I_0 = 2.5$ GW/cm^2 (blue lines), 5 GW/cm^2 (orange lines), 10 GW/cm^2 (yellow lines), 22 GW/cm^2 (purple lines), 50 GW/cm^2 (green lines) and 75 GW/cm^2 (black lines). The solid lines and the dashed lines represent the extracted temperatures from the solution of the BE [Eq. (3) in Ref. [12]] and the effective temperatures obtained from the TTM (see Appendix B), respectively. The inset shows the decay rate of T_e as a function of the electron temperature (the phonon temperature is set to be 300 K). (b) The (instantaneous) chemical potential (see details in Ref. [12]) as a function of time for pulse intensity $I_0 = 50$ GW/cm^2. (c) The electron distribution (solid lines) as a function of energy at $t = -220$ fs (blue), $t = -55$ fs (red), and $t = 110$ fs (green) for the case of $I_0 = 50$ GW/cm^2, the same as (b). The dashed lines are the thermal distributions at the extracted electron temperatures and the vertical dashed lines represent the chemical potential corresponding to each of these lines; a slightly imperfect match is seen only at the early stages.

40-nm thin film of ITO deposited on glass and covered by a 100-nm gold film, see Fig. 1. The real part of the ITO permittivity at 300 K is near zero at approximately 225 THz. This configuration is effective mainly because of the low-wave impedance of the incoming pulse, which enables good light penetration into the ITO.

The electric fields of the incident pump and probe pulses are written as the product of an envelope and a carrier wave. $\vec{\mathcal{E}}_{\text{pump}}(t) = (\mathbf{E}_{\text{pump},0}/2)e^{-2\ln 2(t/\tau_{\text{pump}})^2}e^{-i\omega_{\text{pump}}t} + \text{c.c.}$ and $\vec{\mathcal{E}}_{\text{probe}}(t;\tau) = (\mathbf{E}_{\text{probe},0}/2)e^{-2\ln 2((t-\tau)/\tau_{\text{probe}})^2}e^{-i\omega_{\text{probe}}(t-\tau)}$
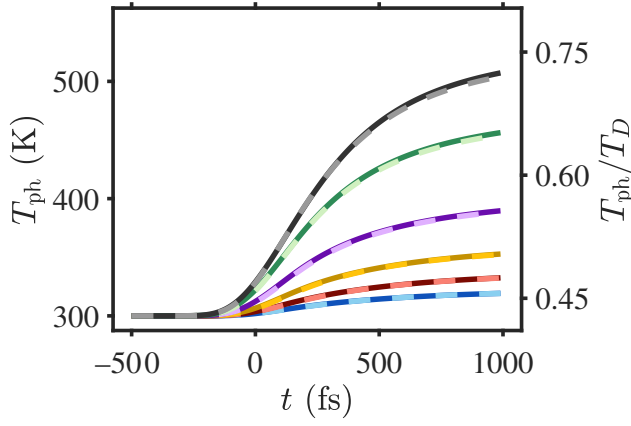
FIG. 3.   The same as Fig. 2(a) but for the phonon temperature $T_{\rm ph}$ (normalization to the Debye temperature on right axis).

+ c.c., where $\mathbf{E}_{\rm pump,0}$ is related to the peak intensity by $I_0 = |\mathbf{E}_{\rm pump,0}|^2/2\sqrt{\dfrac{\varepsilon_0}{\mu_0}}$, and the envelope is taken to be Gaussian [32]. Here $\tau_{\rm pump} = \tau_{\rm probe} = 220$ fs are the pulse durations, $\omega_{\rm pump}/2\pi = \omega_{\rm probe}/2\pi = 230$ THz are the carrier frequencies, τ is the time delay between the arrival of pump and probe pulses, and $\mathbf{E}_{\rm probe,0}$ is the probe-pulse peak field, which is assumed to be weak enough so that it does not affect the electron dynamics in the ITO. Both pump and probe pulses are set to be p polarized to satisfy the ENZ resonance condition. In the self-consistent calculation, we update the local field by solving Eq. (4) via the transfer-matrix method [8,33,34]. Since the thickness of the ITO layer is much smaller than the pump wavelength, the absorbed power density is almost uniform and thus can be approximated by its spatial average for simplicity. Moreover, we assume that the incident pulse has a large spot size so that the nonuniformity of T_e and $T_{\rm ph}$ in the transverse direction can be neglected. The initial temperature of the electrons and phonons are set to be room temperature $T_0 = 300$ K.

We show below results for pump peak intensities of $I_0 = 5$ to 75 GW/cm². For simplicity, we ignore any nonlinearities in the glass and gold layers. This is justified in light of the extremely strong nonlinearity of ITO. For even higher intensities, substantial amounts of electrons would accumulate at the top edge of the conduction band. Resolving this requires one to account for spontaneous and stimulated emission of photons, photoemission of electrons, and even for higher-order nonparabolicity of the band structure, multiphoton processes etc. The study of these effects is left to future studies. We also compare our electron dynamics model with the TTM, see details in Appendix B. This provides a better understanding of the effects of the nonthermal electron distribution on the ITO nonlinearity.
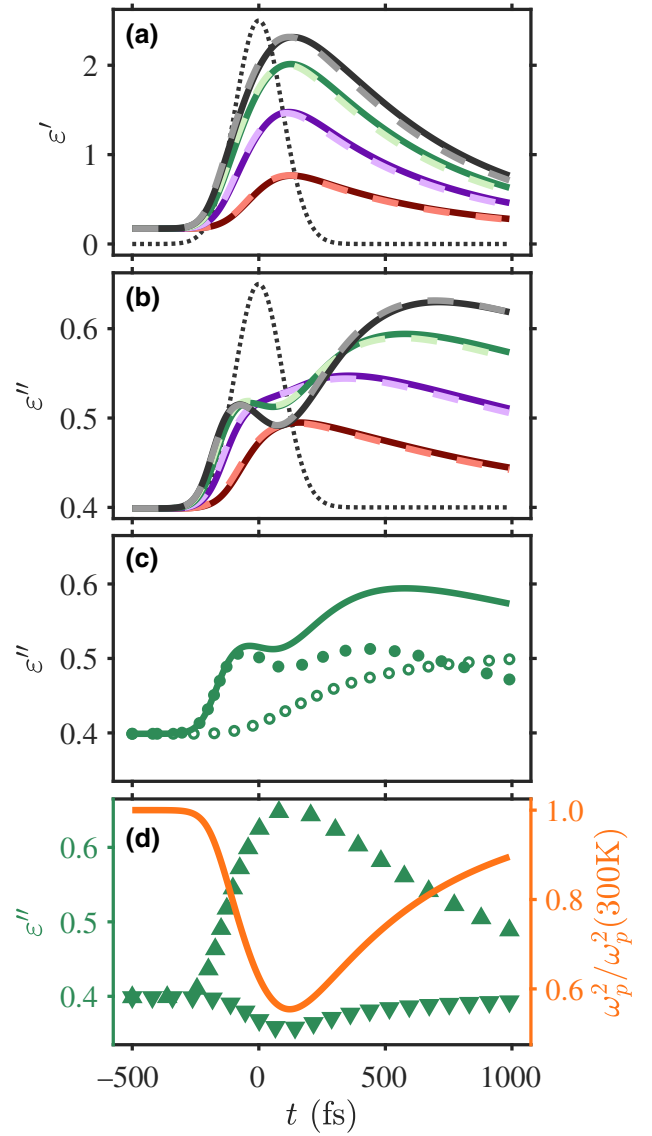

FIG. 4.   (a) The real part and (b) the imaginary part of the permittivity, Eq. (9), as a function of the time delay for the pulse intensity $I_0 = 5$ GW/cm², 22 GW/cm², 50 GW/cm², and 75 GW/cm² (the color used is the same as Fig. 2) at the carrier frequency. (c) The separation of the contributions to ε'' due to the change of $\eta_{e\text{-ph}}$ (open dots); and T_e (filled dots) for $I_0 = 50$ GW/cm². The solid line is the same as in (b). (d) The separation of the effects of T_e on ε'' via the change of $\eta_{e\text{-}e}$ (up-pointing triangles) and ω_p^2 (down-pointing triangles). The orange solid line shows the change of ω_p^2 (normalized to its value at 300 K) as a function of the time delay. The thin black dotted lines in (a),(b) represent the pump-pulse intensity profile.

A. Temperature dynamics

For the purpose of characterizing the solution of the ANTH$\mathscr{E}$M quantitatively, we extract the electron temperature (denoted as *extracted* T_e) from the electron distribution based on the total energy of the electron system. The total energy of the electron system $\mathcal{U}$ is the first moment of
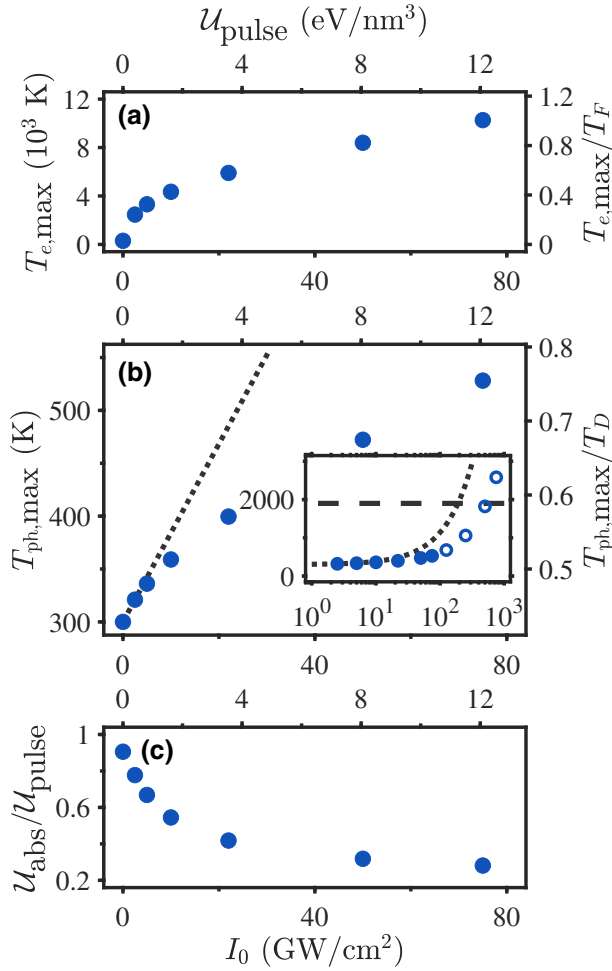
FIG. 5. (a) The maximum of the electron temperature (normalized to the Fermi temperature on right axis) as a function of the pulse peak intensity [extracted from Fig. 2(a)]. (b) The same as (a) but for the phonon temperature $T_{\rm ph}$ (normalization to the Debye temperature on right axis). The (filled) dots represent the data obtained from the simulation and the dotted black line represents the maximum phonon temperature obtained from an intensity-independent permittivity approximation. The inset of (b) shows the same data on a semilogx scale and is extended to 750 GW/cm². The open dots represent the estimated upper bound of the maximum $T_{\rm ph}$ for I_0 up to 750 GW/cm² assuming that the absorptivity is constant for $I_0 > 75$ GW/cm². The dashed black line represents the damage threshold temperature (1900 K) of ITO reported in Ref. [8]. (c) The total absorbed pump-pulse energy (normalized to the pump-pulse energy) as a function of the pulse peak intensity (bottom x axis) and pump-pulse energy density (top x axis).

the electron distribution, i.e., $\mathcal{U} = \int \mathcal{E}\rho_e(\mathcal{E})f(\mathcal{E})d\mathcal{E}$. The electron temperature associated with an electron distribution $f(\mathcal{E})$ is then determined by

$$\int \mathcal{E}\rho_e(\mathcal{E})f^T(\mathcal{E}, \mu(T_e), T_e)d\mathcal{E} = \int \mathcal{E}\rho_e(\mathcal{E})f(\mathcal{E})d\mathcal{E},$$

where $\mu(T_e)$ is the chemical potential given by Eq. (B5), see details in Ref. [12]. We plot the *extracted* electron temperature along with the *effective* electron temperature obtained from the TTM as a function of time for increasing illumination intensities. Figure 2(a) shows the dynamics of a rapid (pulse duration-limited) increase followed by electron cooling familiar from noble metals [20]. However, Fig. 2(a) also shows that, unlike noble metals, the electron subsystem reaches extremely high electron temperatures (in particular, 10^4 K), i.e., a considerable fraction of the Fermi temperature (approximately T_F). Notably, since the pump-pulse duration is much longer than the *e-e* relaxation time $\tau_{e\text{-}e}$ [a few tens of fs, see Fig. 9(a)], the effective T_e shows an excellent agreement with the dynamics of the extracted T_e. As a result, the T_e dynamics can be understood by energy balance [Eq. (B12a)], where the electron-temperature decay rate dT_e/dt is related to the electron-heat capacity C_e, the electron-phonon energy-coupling coefficient $G_{e\text{-}ph}$ and to the temperature difference between electrons and phonons [35] by $dT_e/dt = -G_{e\text{-}ph}(T_e)(T_e - T_{\rm ph})/C_e(T_e)$. The inset of Fig. 2(a) shows that the electron-temperature decay rate is faster for higher T_e, or more specifically, the slopes of the $T_e(t)$ curves become less negative as t increases (T_e decreases). This slower decay originates from the proportionality to the temperature difference between electrons and phonons and the simultaneous mere sublinear increase of C_e with T_e (see Ref. [12]). Figure 2(a) also shows that the slope of $T_e(t)$ is more negative for higher I_0 (higher T_e).

The drastic increase of T_e for $I_0 \geq 50$ GW/cm² reveals a rather surprising result—the effective chemical potential μ can become negative. Specifically, Fig. 2(b) plots μ as a function of time for the case of $I_0 = 50$ GW/cm², showing that the chemical potential goes below zero after the pulse peaks at 0 fs, and again recovers to positive values as T_e decays. Consequently, the electron distribution is initially flatter, as seen in Fig. 2(c), and eventually becomes a distribution that is analogous to the electron distribution in a semiconductor [26].

The cooling of the electrons results in an increase of $T_{\rm ph}$, see Fig. 3. Figure 3 also shows excellent agreement between the $T_{\rm ph}$ dynamics in the ANTH$\mathscr{E}$M and the TTM [Eq. (B12)]. Moreover, the phonon-temperature dynamics show a decreasing growth rate with time.

B. Permittivity dynamics (nonlinear optical response)

In Fig. 4 we plot the dynamics of the ITO permittivity based on the solution of the ANTH$\mathscr{E}$M. The result shows a remarkable match to a permittivity calculation that relies on pure thermal electron distributions with the effective T_e. This match enables a simple interpretation of the permittivity in terms of the dynamics of the temperatures shown in Figs. 2(a) and 2(b). This excellent match is remarkable since it does not require any of the phenomenological

corrections introduced in Ref. [9]. In the latter, the permittivity calculations matched the experimental reflectance and transmittance [10] only if the damping coefficient η is assumed to increase linearly with T_e, the e-ph coupling coefficient is reduced by a factor of 8, and the e-ph energy transfer from the nonthermal part of the electron distribution is assumed to be negligible. The self-consistent treatment presented here not only naturally connects the damping η with T_e and T_{ph}, but also explains the experimentally observed increase of the imaginary part of the permittivity induced by the illumination [14].

The most striking aspect of the dynamics is the drastic change in the real part ε' (a 12-fold increase for $I_0 = 50$ GW/cm^2), which is a result of the initial proximity to the ENZ point [7,8,10,36]. One can see that the dynamics of ε' shares similar features with the time evolution of T_e, [compare Figs. 2(a) and 4(a)]. This is caused by the (nearly) linear decrease of the effective plasma frequency with T_e due to the nonparabolicity, which has been explained in many previous studies [3,4,7,37]. Therefore, by Fig. 2(a), the temporal decay of the ITO permittivity is faster for higher effective T_e. This indicates that the permittivity dynamics shows the opposite behavior to the experimentally observed slowing of the reflection decay at high intensities in Ref. [8]. We return to this issue in Sec. III D. The positive correlation between ε' and the effective T_e also explains why the maximum change in ε' occurs later than the peak of the pump pulse. This is because in the early stages the change of effective T_e is approximately proportional to the total absorbed energy density, i.e., $\Delta \varepsilon'(t) \sim \Delta T_e(t) \sim \int_{-\infty}^{t} p_{abs}(t')dt'$.

In that regard, although ITO is a Drude material, the (larger than realized before) temperature-induced change in the real part of the permittivity makes the ENZ nonlinear optical response of ITO qualitatively different from that of noble metals; indeed, in the latter, the light-induced changes to the imaginary part dominate over the changes to the real part [23,38]. Moreover, at the ENZ point, not only the field enhancement and absorption are strong, but also the relative change ε' ($= \Delta\varepsilon'/\varepsilon'$) and thus the nonlinear optical response is maximized.

By contrast, ε'' shows a more complicated dynamics because it is proportional to the product of ω_p^2 [Eq. (10)] and the damping term η, and thus depends on both T_e and T_{ph}. To see this, in Fig. 4(c), we separate the contribution due to the change of T_e (by setting $T_{ph} = 300$ K) and of T_{ph} (setting $T_e = 300$ K). Since T_{ph} affects $\tau_{e\text{-}ph}^{-1}$ linearly [12] but does not affect $\tau_{e\text{-}e}^{-1}$ nor ω_p [see Eq. (10)], the contribution due to T_{ph} results in a slow increase of ε'' [see open dots in Fig. 4(c)]. Figure 4(c) also shows that the complicated dynamics of ε'' originates from the dependence on T_e [see solid dots in Fig. 4(c)]. In particular, when T_e increases within the duration of -200 fs $< t < 0$, $\tau_{e\text{-}e}^{-1}$ increases but
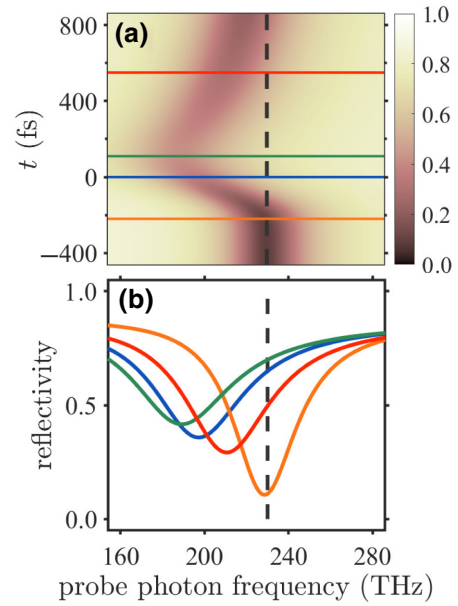

FIG. 6. (a) The map of the reflectivity (the absolute square of the Fresnel reflection coefficient) for the pump peak intensities of 22 GW/cm^2, respectively. (b) The reflectivity of the probe pulse at time-delay sections $\tau = -220$ (orange), 0 (blue), 110 (green), and 550 fs (red) in (a),(c), respectively. The black dashed line represents the center frequency (230 THz) of the probe pulse.

ω_p decreases, see their opposite effects on the ε'' dynamics separated in Fig. 4(d). Thus, the T_e dependence of ε'' can be roughly expressed as $\varepsilon'' \sim T_e^2(1 - \alpha T_e)$ [39], indicating that ε'' increases for modest electron temperatures but decreases at higher ones. Therefore, when the pump peak intensity is relatively high [> 50 GW/cm^2 in Figs. 4(b) and 4(c)], the standard ε'' dynamics of a rapid (pulse duration-limited) increase followed by electron cooling and phonon heating is accompanied by an additional drop once high electron temperatures are reached [see the green solid lines in Figs. 4(b) and 4(c)]. That drop in ε'' is more pronounced as the illumination intensity is increased, compare the purple and green lines in Fig. 4(b). Of note, the increase of the imaginary part of the ITO permittivity with time (and intensity) further indicates that the physics underlying the thermo-optic response of LEDD materials is similar to that of Drude metals (see Refs. [21,22], and is the opposite to that of saturable absorbers (for which the imaginary part of the permittivity decreases upon illumination).

C. Intensity dependence and energy partition

One of the direct consequences of the drastic change in ε' induced by the strong illumination is the sublinear growth of the maximum of the electron and phonon temperatures with the pump peak intensity, as shown in Figs. 5(a) and 5(b). The maximal values of T_{ph}, although can be estimated by extrapolating the T_{ph} dynamics in

Fig. 2(b), are more easily deduced from the total absorbed energy density $\left(\mathcal{U}_{abs}(I_0) = \int p_{abs}(t, I_0) dt \right)$ from the self-consistent simulations via $T_{ph,\max}(I_0) = T_0 + \mathcal{U}_{abs}(I_0)/C_{ph}$. This is based on the fact that almost all the absorbed energy transfers to the phonon subsystem before leaking out to the surrounding [40]. The sublinear growth of the maximum phonon temperatures with the pump peak intensity can then be understood by the absorptivity (the ratio of the total energy absorbed by ITO and the incident pump-pulse energy, denoted as $\mathcal{U}_{pulse}$) as shown in Fig. 5(c). It shows that at low illumination intensities, for which the permittivity is close to the ENZ point (see Fig. 4), most energy is absorbed (ENZ resonance). However, as the illumination intensity grows and the real part of the ITO permittivity drastically increases [see Fig. 4(a)], the resulting shift of the ENZ point causes the absorptivity to drop rapidly and the fraction of the incident energy reflected to grow. This effect can be captured only by a self-consistent simulation. This confirms that the decrease of the absorptivity does not originate from saturable absorption.

In Fig. 5(b), we further extrapolate the data obtained from the simulations up to 750 GW/cm^2 (the open dots) by assuming that the absorptivity remains the same for $I_0 > 75$ GW/cm^2, namely, $T_{ph,\max}(I_0 > 75$ GW/cm$^2) = T_0 + [\mathcal{U}_{abs}(I_0 = 75$ GW/cm$^2)/C_{ph}][I_0/(75$ GW/cm$^2)]$. The open dots in Fig. 5(b) indicates that the resulting (somewhat

overestimated) T_{ph} reaches the melting point of ITO (approximately 1900 K) for $I_0 > 500$ GW/cm^2. This explains the high damage threshold of ITO observed in Ref. [8].

Finally, we note that the sublinearity of maximal T_e versus I_0 is more pronounced than that of T_{ph}. This is because, in addition to the decrease of absorptivity with I_0, the electron heat capacity increases with T_e such that more energy is required to change T_e at higher T_e, yielding a smaller growth of maximal T_e with the peak intensity.

D. Probe-pulse dynamics

In order to connect the ANTH$\mathscr{E}$M to the experimental data [8], we now calculate the temporal reflectivity and the total reflection of the probe pulse by the sample in the presence of the pump pulse. Since the probe pulse satisfies the adiabatic condition (see Sec. II B), Ampère's law and the Maxwell-Faraday equation for the probe pulse become

$$\begin{cases} \nabla \times \mathbf{H}_{probe}(\mathbf{r}, t; \tau) = -i\omega_{probe}\varepsilon_0 \\ \qquad\qquad \varepsilon(t; \omega_{pump}, \omega_{probe})\mathbf{E}_{probe}(\mathbf{r}, t; \tau) \\ \nabla \times \mathbf{E}_{probe}(\mathbf{r}, t; \tau) = -i\omega_{probe}\mu_0\mathbf{H}_{probe}(\mathbf{r}, t; \tau) \end{cases}, \quad (11)$$

where $\varepsilon(t; \omega_{pump}, \omega_{probe})$ is the time-dependent permittivity for the probe pulse, namely,

$$\varepsilon(t; \omega_{pump}, \omega_{probe}) = \varepsilon_\infty + \lim_{\mathbf{q} \to 0} \frac{2e^2}{\varepsilon_0 q^2} \int \frac{d^3k}{(2\pi)^3} \frac{f_{\mathbf{k+q}}(t; \omega_{pump}) - f_{\mathbf{k}}(t; \omega_{pump})}{\mathcal{E}_{\mathbf{k+q}} - \mathcal{E}_{\mathbf{k}} - \hbar\omega_{probe} - i\hbar(\eta_{\mathbf{k+q}}(t; \omega_{pump}) + \eta_{\mathbf{k}}(t; \omega_{pump}))/2}. \quad (12)$$

Since the intensity of the probe pulse is weak enough so that it does not induce significant changes to the electron distribution, the electron distribution in Eq. (12) is the self-consistent solution obtained from Eqs. (1), (4), and (5). Again, the time variable t in Eqs. (11) and (12) is correlated with the pump-pulse envelope but is not the Fourier conjugate of any photon frequency (see Sec. II B). Based on the adiabaticity, the temporal profile of the reflected probe-pulse envelope $\mathbf{E}_{probe,ref}(t; \tau)$ is thus equal to the product of the Fresnel reflection coefficient and the incident probe-pulse envelope, i.e., $\mathbf{E}_{probe,ref}(t; \tau) = r(t; \omega_{probe})(\mathbf{E}_{probe,0}/2)e^{-2\ln 2((t-\tau)/\tau_{probe})^2}$. Notice that the Fresnel reflection coefficient is *independent* of the time-delay parameter since it originates from the (pump-induced) electron dynamics; its absolute square (i.e., the reflectivity) plotted in Fig. 6 is thus different from the spectrum of the reflected probe pulse (see below) [41]. Figure 6 shows that the drastic change of the real part of the permittivity induced by the pump

pulse results in a shift of the ENZ resonance frequency and that this shift increases with the pump peak intensity.

The reflection of the probe pulse is then the ratio of the total energy of the reflected pulse and that of the incident probe pulse, $R(\tau) = \int |\mathbf{E}_{probe,ref}(t; \tau)|^2 dt / \int |\mathbf{E}_{probe,inc}(t; \tau)|^2 dt$; it is, therefore, a function of the time delay. Figure 7(a) shows that the reflection of the probe pulse gradually increases with the pump peak intensity, even though at a decreasing rate. In addition, the time in which the change of the reflection reduces from its maximum to half of this maximum is longer for higher pump peak intensities, see Fig. 7(b). This finding matches the observations in Ref. [8]. The reason for that is that when the pump peak intensity is high enough, the induced frequency shift of the ENZ resonance is larger than its resonance width [see Fig. 6(b)] so that the dependence of the reflection change on the pump peak intensity becomes sublinear.
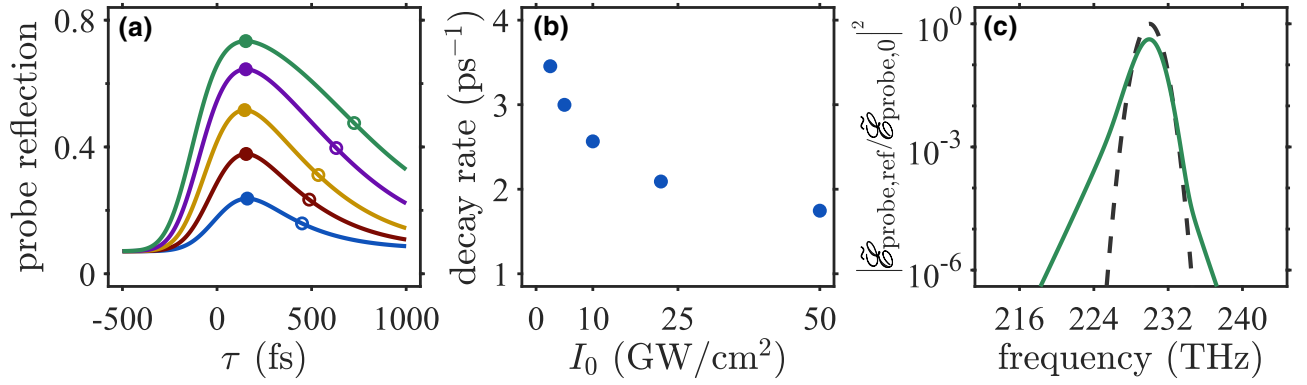
FIG. 7. (a) The reflection of the probe pulse (the ratio of the total energy of the reflected pulse and that of the incident probe pulse) as a function of the pump-probe delay for different pump-pulse peak intensities (the same colors are used as in Fig. 4). The filled and open dots label the maximum and half-maximum of the reflection, respectively. (b) The decay rate (the inverse of the time period corresponding to the change of the reflection from its maximum to half this maximum) as a function of the pulse peak intensity. (c) The spectrum of the reflected probe pulse at the time delay of -90 fs (green solid line) and -500 fs (black dashed line) for the pump-pulse peak intensity of 50 GW/cm^2.
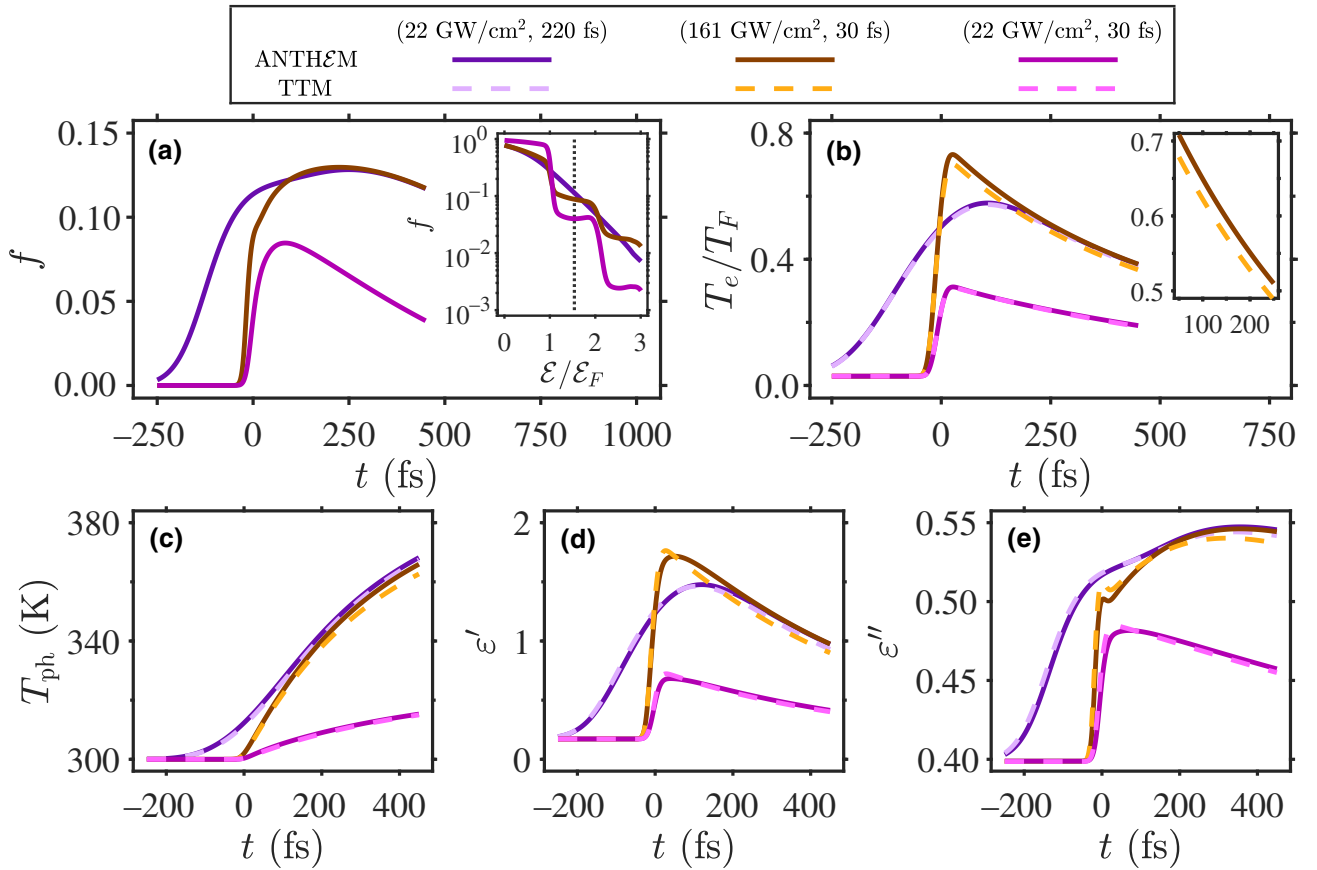

FIG. 8. The time dependence of (a) the electron distribution at $\mathcal{E} = \mathcal{E}_F + \hbar\omega_{\text{pump}}/2$, (b) the electron temperature (normalized to the Fermi temperature), (c) the phonon temperature (normalized to the Debye temperature), (d) the real part, and (e) the imaginary part of the ITO permittivity. The solid and dashed lines in (b),(c) are obtained from the ANTH$\mathscr{E}$M and from the TTM, respectively. In (d),(e), the solid and dashed lines are calculated from Eq. (5) using the electron distribution obtained from the ANTH$\mathscr{E}$M and the thermal distribution with the effective T_e, respectively. The purple, brown (or orange), and magenta lines, respectively, represent the cases of 22 GW/cm^2 and 220 fs, 161 GW/cm^2 and 30 fs, and 22 GW/cm^2 and 30 fs. The inset in (a) shows the electron distribution at $t = 0$ fs. The black dotted line labels the energy $\mathcal{E} = \mathcal{E}_F + \hbar\omega_{\text{pump}}/2$. The inset in (b) is the enlargement at $t = 50 - 250$ fs, showing the difference in T_e between the ANTH$\mathscr{E}$M and TTM (approximately 200 K) more clearly.

The lack of time translational invariance prevents us from writing the spectrum of the reflected probe pulse as a product of the Fresnel reflection coefficient and the Fourier transform of the incident probe pulse. Therefore, we calculate the spectrum of the reflected probe pulse directly from its Fourier transform, namely, $\vec{\tilde{\mathscr{E}}}_{\text{probe,ref}}(\omega; \tau) = \int \vec{\mathscr{E}}_{\text{probe,ref}}(t; \tau)e^{i\omega t}dt$. Here, $\vec{\mathscr{E}}_{\text{probe,ref}}$ $(t; \tau) = \mathbf{E}_{\text{probe,ref}}(t; \tau)e^{-i\omega_{\text{probe}}t} + \text{c.c.} = r(t; \omega_{\text{probe}})$ $(\mathbf{E}_{\text{probe},0}/2)e^{-2\ln 2((t-\tau)/\tau_{\text{probe}})^2 - i\omega_{\text{probe}}(t-\tau)} + \text{c.c.}$ is the electric field of the reflected probe pulse. The adiabaticity ensures that the Fresnel coefficient, evaluated only at the center frequency of the probe pulse [i.e., $r(t, \omega_{\text{probe}})$, where ω_{probe} corresponds to the black dashed line in Fig. 6(a)], is enough to calculate $\vec{\tilde{\mathscr{E}}}_{\text{probe,ref}}(\omega; \tau)$. Hence, the significant ENZ resonance frequency shift does not appear in the spectrum of the reflected probe pulse, see Fig. 7(c). Moreover, due to the transient change of the reflectivity induced by the pump pulse new frequencies are generated beyond the bandwidth of the spectrum of the incident probe pulse, as shown in Fig. 7(c). Since the rising time of ε' [see Fig. 4(a)] is long compared to the periodicity of the probe pulse $2\pi/\omega_{\text{probe}}$, the widening of the reflected probe-[pulse spectrum is rather weak (see the comparison between green solid and black dashed lines in Fig. 7(c) at the respective FWHM). These results, therefore, agree with the observations in Ref. [8].

E. The response to shorter pulses

Since the duration of the pump pulses considered in previous subsections and in the experiments [2–5,7–9] are much longer than the *e-e* relaxation time, the dynamics of the temperatures and of the permittivity can be well approximated by the TTM, see also the interpretation of the experimental results [10] in Ref. [9]. To go beyond that description, we consider a pump pulse with a duration of 30 fs, which is comparable with the *e-e* relaxation time but still satisfies the condition for the adiabatic approximation, so that we can employ the model described in Sec. II to calculate the electronic and optical responses. To make a fair comparison, the peak intensity of the 30-fs pulse is chosen to be 161 GW/cm^2 so that the pulse energy is the same as the 220-fs pulse with $I_0 = 22$ GW/cm^2 used so far. We further compare the results based on ANTH$\mathscr{E}$M with the TTM as well as on the extended TTM (ETTM), see details in Appendix B. The comparison above therefore allows us to pinpoint the role of the nonthermal electrons on the nonlinear response of ITO.

Figure 8 shows that the dynamics of the electron distribution, extracted electron temperature, and permittivity for the case of the 30-fs pulse with peak intensity 161 GW/cm^2 (brown solid lines) are qualitatively the same as the case of 220-fs pulse with peak intensity 22 GW/cm^2 (purple solid lines) studied in Figs. 2 and 4, except for moderate

quantitative differences in the early stages. Specifically, after the pump pulse hits the sample, the extracted electron temperature and the permittivity for the case of 30-fs pulse with $I_0 = 161$ GW/cm^2 increase rapidly on the time scale of the pulse duration, much faster than that for the case of 220-fs pulse with $I_0 = 22$ GW/cm^2 [see Figs. 8(a), 8(b), 8(d), and 8(e)], resulting in a similar probe-pulse reflection decay and a much wider reflected pulse spectrum (see Fig. 11 in Appendix D). In addition, comparing with the case of 220-fs pulse with $I_0 = 22$ GW/cm^2, during the much shorter T_e rise time there is much less energy transfer from the electron to the phonon subsystem, resulting in a higher maximal value of the electron temperature [Fig. 8(b)], and thus, in a larger change of ε' [Fig. 8(d)], and a more pronounced drop in the dynamics of ε'' [Fig. 8(e)].

Moreover, Figs. 8(b) and 8(c) show that for the case of 30-fs pulse with $I_0 = 161$ GW/cm^2, the effective T_e is around 200 K lower and T_{ph} is a few K lower than the results of the ANTH$\mathscr{E}$M. This occurs because comparing with the ANTH$\mathscr{E}$M, the instantaneous thermalization assumption in the TTM causes the change of ε' to be slightly larger [see Figs. 8(d)], resulting in a somewhat smaller absorption. This indicates that the instantaneous thermalization assumption in TTM becomes less valid for shorter pulses.

To further understand the role of the nonthermal electron distribution on the nonlinear optical response of ITO, we compare the ANTH$\mathscr{E}$M with the ETTM (see details in Appendix C). In contrast to the TTM, the ETTM accounts for the finite *e-e* relaxation time and thus allows the electron subsystem to be nonthermal. In this model, the electron subsystem is described by an *instantaneous* temperature that represents only the thermal part of the electron distribution. This approach is essential when the *e-e* thermalization time is comparable with the *e*-ph relaxation time, see Refs. [42,43]. Figure 10 shows that the results of the ETTM are in remarkable agreement with the ANTH$\mathscr{E}$M except that the rise time of the instantaneous T_e is controlled by the *e-e* relaxation time instead of the pulse duration and thus is longer than that of the extracted T_e (see discussion in Appendix C). Therefore, the response time of ε' is controlled by the pulse duration due to the account for the nonthermal distribution in the permittivity calculation. Moreover, if the nonthermal contribution is neglected when calculating the permittivity, the response time of ε' will be controlled by the *e-e* relaxation time, see Fig. 10(b). As a result, the absorption [Fig. 10(c)] and thus the maximum value of the instantaneous T_e [Fig. 10(a)] are much higher compared with the results accounting for the nonthermal distribution in the permittivity calculation. This implies that the instantaneous rise time of ε' is due to the change of the population, as claimed in Ref. [44]. Therefore, the account for the nonthermal electron distribution is an absolute requirement

for the accurate description of the nonlinear response of ITO on the few fs timescale.

From the comparison above one can also see that comparing with the TTM, although the ETTM is more accurate, the result of the ETTM is only slightly better. This is quite different from noble metals [42] and graphite [43]. The main reason for that is the short *e-e* relaxation time in ITO. This justifies the claim for the superiority of the TTM over the ETTM in Ref. [9].

IV. DISCUSSION

Our model is based on a simpler model that we developed for (parabolic band) metals [15,25] and semiconductors [26]. In that simpler model, due to weak illumination intensity, we did not incorporate a self-consistent solution of field, distribution, and permittivity. Our model can also be generalized to two-dimensional materials (for example, graphene, the surface of three-dimensional topological insulators [43,45,46]) with proper modifications of the *e*-ph and *e-e* interaction terms. This work is a starting point for the study of the nonlinear optical response of ITO at shorter wavelengths (e.g., Ref. [3]), nonlinear effects requiring very high nonlinearity such as bistability, and to the study of the response at higher intensities and shorter pulses, and the formulation can also be directly implemented to study the nonlinear optical response of other transparent conductive oxides such as TiN, ZrN, [47] etc. Our present work, along with the previous results from Ref. [12], provides a thorough understanding of the electron dynamics and the associated nonlinear response. This can further aid in understanding the benefits and limitations of transparent conductive oxides for a list of applications [48,49] such as pulse shaping and optical switching, efficient frequency conversion, Terahertz emission, etc. ITO being a CMOS-compatible material, all these applications are directly relevant to our findings. Moreover, the improved understanding of the heat dynamics in ITO systems is crucial to optical communication systems concerning the development of modern data centers.

The nonthermal (and the simpler thermal) permittivity model described above are frequently replaced by a simpler approach—based on assigning a local-like $\chi^{(3)}$ (or n_2) value for the ITO nonlinearity, see, e.g., Refs. [10,50], or later in Refs. [7,9,14]). However, it is well known that for absorptive materials such as metals (and hence, LEDDs), the response is strongly nonlocal in time, such that the fitted values strongly depend on the pulse duration, see, e.g., Ref. [51,52]). As a particular example, Fig. 8 also shows that if the pump peak intensity is reduced to 22 GW/cm^2 (by a factor approximately 8) but the pulse duration remains 30 fs, the changes of the extracted T_e and ε' are only reduced by 50%. The inclusion of higher-order nonlinear coefficients cannot

overcome this difficulty. Moreover, the local nonlinear susceptibility values cannot capture the asymmetric temporal response [namely, the smeared nonlinearity turn off, see Figs. 8(d) and 8(e)]. This prevents one from being able to directly connect the permittivity to the local field intensity, or from being able to characterize the ultrafast nonlinear response of ITO by Kerr-like nonlinearity with an intensity-dependent refractive index, e.g., in Refs. [7,10,53].

V. CONCLUSIONS

We present a fully nonthermal nonphenomenological microscopic model for the electron permittivity of LEDD materials. Our work resolves several arguments and explains several experimental observations reported in the LEDD material literature that is so far not understood, namely, the significance of momentum conservation in *e*-ph interactions, the scaling of the collision rate with T_e, and the negligible energy transfer from nonthermal electrons to phonons. Our model also establishes the instantaneous nature of the pump-induced permittivity changes, and reveals the possibility of the chemical potential to transiently decrease dramatically, even to negative values. From our analysis we conclude that the nature of the nonlinear optical response of ITO (and LEDD materials) is not polynomial nor saturable, but rather a complex temporally local nonequilibrium response that can be approximated by thermal models only for sufficiently long pulses.

APPENDIX A: THE ADIABATIC APPROXIMATION

To solve Maxwell equations for systems whose material properties are time dependent, we first derive the polarization density by solving the density-matrix equations in the time domain and take the trace of the product of the transition dipole moment with the density matrix (see, e.g., Ref. [16]). The electric displacement $\vec{\mathscr{D}}(t)$ is then related to the local electric field $\vec{\mathscr{E}}(t)$ by [54,55]

$$\vec{\mathscr{D}}(t) = \varepsilon_0 \varepsilon_\infty \vec{\mathscr{E}}(t) + \int_{-\infty}^{t} R(t,t')\vec{\mathscr{E}}(t')dt', \qquad (A1)$$

where $R(t,t')$ is the memory function and is related to the electron distribution by

$$R(t,t') = \lim_{\mathbf{q}\to 0} \frac{2e^2}{i\hbar q^2} \int \frac{d^3k}{(2\pi)^3} (f_{\mathbf{k}+\mathbf{q}}(t') - f_{\mathbf{k}}(t'))$$
$$e^{-[(\eta_{\mathbf{k}+\mathbf{q}}(f(t')) + \eta_{\mathbf{k}}(f(t')))/2 + i(\mathcal{E}_{\mathbf{k}+\mathbf{q}} - \mathcal{E}_{\mathbf{k}})/\hbar](t-t')}. \quad (A2)$$

If the damping rate $\eta_{\mathbf{k}}$ is much faster than the rate of change of the electron distribution, only the electron dynamics at t' nearby t contributes to the integral in Eq. (A1). Since the change rate of the electron distribution

is caused by absorption via the electron excitation term, which is incoherent in nature (i.e., it depends on $|\mathscr{E}|^2$), the condition above is satisfied when $\left|(df/dt)_{\mathrm{exc}}\right| < \eta/\hbar$. This requires the local field to be smaller than 2.7×10^9 V/m [56]. In this case, one can replace $f_{\mathbf{k}}(t')$ by $f_{\mathbf{k}}(t)$ and $\eta_{\mathbf{k}}(f(t'))$ by $\eta_{\mathbf{k}}(f(t))$, so that we can factor the electron distribution out of the integral. We further assume that the pulse duration is much longer than the periodicity of the carrier wave ($\tau_{\mathrm{pump}} \gg 2\pi/\omega_{\mathrm{pump}}$) and is also much longer than the damping time ($\tau_{\mathrm{pump}} \gg 1/\eta$). The condition $\tau_{\mathrm{pump}} \gg 2\pi/\omega_{\mathrm{pump}}$ allows us to write the electric field (and the electric displacement) as a product of slowly varying envelope [$\mathbf{E}(t')$ or $\mathbf{D}(t')$] and a rapidly varying phase factor with the carrier frequency, i.e., $\vec{\mathscr{E}}(t') = \mathbf{E}(t')e^{-i\omega_{\mathrm{pump}}t'} + \mathrm{c.c.}$ ($\vec{\mathscr{D}}(t') = \mathbf{D}(t')e^{-i\omega_{\mathrm{pump}}t'} + \mathrm{c.c.}$). After substituting $\vec{\mathscr{E}}(t')$ back to Eq. (A1), the condition $\tau_{\mathrm{pump}} \gg 1/\eta$ allows us to replace $\mathbf{E}(t')$ by $\mathbf{E}(t)$ so that the envelope function $\mathbf{E}(t)$ can be factored out of the t'-integral. Then, Eq. (A1) becomes

$$\mathbf{D}(t) = \varepsilon_0\varepsilon_\infty\mathbf{E}(t) + \lim_{\mathbf{q}\to 0}\int \frac{d^3k}{(2\pi)^3}(f_{\mathbf{k}+\mathbf{q}}(t) - f_{\mathbf{k}}(t))\mathbf{E}(t)\frac{2e^2}{i\hbar q^2}$$

$$\times \int_{-\infty}^t dt' e^{-((\eta_{\mathbf{k}+\mathbf{q}}(t)+\eta_{\mathbf{k}}(t))/2 + i(\mathcal{E}_{\mathbf{k}+\mathbf{q}}-\mathcal{E}_{\mathbf{k}})/\hbar)(t-t')}$$

$$e^{i\omega_{\mathrm{pump}}(t-t')}. \quad \text{(A3)}$$

After performing the t'-integral, Eq. (A3) can be reorganized into the form $\mathbf{D}(t) = \varepsilon_0\varepsilon(t;\omega_{\mathrm{pump}})\mathbf{E}(t)$, where $\varepsilon(t;\omega_{\mathrm{pump}})$ is given by Eq. (5). Comparing with Eqs. (A1) and (A2), the approximations (A3) and (5) state that the time-dependent permittivity changes with electron distribution adiabatically.

When substituting the electric displacement field into Ampère's law, the time derivative of the electric displacement field becomes

$$\frac{\partial\vec{\mathscr{D}}(t)}{\partial t} = \varepsilon_0\left[\frac{\partial\varepsilon(t;\omega_{\mathrm{pump}})}{\partial t}\mathbf{E}(t) + \varepsilon(t;\omega_{\mathrm{pump}})\frac{\partial\mathbf{E}(t)}{\partial t}\right.$$

$$\left. - i\omega_{\mathrm{pump}}\varepsilon(t;\omega_{\mathrm{pump}})\mathbf{E}(t)\right]e^{-i\omega_{\mathrm{pump}}t} + \mathrm{c.c.}$$

Since the oscillation of the carrier wave is much faster than the change rate of the envelope $\mathbf{E}(t)$ and of $\varepsilon(t;\omega_{\mathrm{pump}})$, the time derivative of the electric displacement field can be approximated by $\partial\vec{\mathscr{D}}(t)/\partial t \approx -i\omega_{\mathrm{pump}}\varepsilon_0\varepsilon(t;\omega_{\mathrm{pump}})\mathbf{E}(t)e^{-i\omega_{\mathrm{pump}}t} + \mathrm{c.c.}$ so that Ampère's law becomes $\nabla \times \mathbf{H}(\mathbf{r},t) = -i\omega_{\mathrm{pump}}\varepsilon_0\varepsilon(t;\omega_{\mathrm{pump}})\mathbf{E}(\mathbf{r},t)$. Similarly, the Maxwell-Faraday law becomes $\nabla \times \mathbf{E}(\mathbf{r},t) = -i\omega_{\mathrm{pump}}\mu_0\mathbf{H}(\mathbf{r},t)$, see Eq. (4). The solution of Eq. (4) and the time-dependent permittivity (5) are then used to calculate the power-absorbed density via the Ponyting theorem, see Eq. (6). When the pulse duration

is comparable to the carrier-wave periodicity, one should go beyond the adiabatic approximation and calculate the absorbed power density using Poynting's theorem in the time domain.

APPENDIX B: EXTENDED TWO-TEMPERATURE MODEL AND TWO-TEMPERATURE MODEL

Following the approach in Refs. [15,25], the dynamics of the electron and phonon subsystems can be macroscopically described by deriving the extended two temperature model by coarse graining the Boltzmann equation (1). In this approach, we split the Boltzmann equation (1) for the electron distribution into a thermal and a nonthermal part, $f(\mathcal{E}) = f^T(\mathcal{E}, \mu(T_e), T_e) + f^{\mathrm{NT}}(\mathcal{E})$, and add the energy-balancing equation for the phonon subsystem. We then multiply the Boltzmann equation by the product of the electron energy, $\mathcal{E}$, and the density of electron states, $\rho_e(\mathcal{E})$, and integrate over the electron energy. This results in a pair of equations describing the dynamics of the nonthermal electron distribution f^{NT} and its integral version [i.e., an equation for the nonthermal electron energy $\mathcal{U}^{\mathrm{NT}} = \int \mathcal{E}\rho_e(\mathcal{E})f^{\mathrm{NT}}d\mathcal{E}$],

$$\frac{\partial f^{\mathrm{NT}}}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{\mathrm{exc}} + \left(\frac{\partial f^{\mathrm{NT}}}{\partial t}\right)_{e\text{-}e} + \left(\frac{\partial f^{\mathrm{NT}}}{\partial t}\right)_{e\text{-ph}}, \quad \text{(B1a)}$$

$$\frac{d\mathcal{U}^{\mathrm{NT}}}{dt} = p_{\mathrm{abs}}(t) + \left(\frac{d\mathcal{U}^{\mathrm{NT}}}{dt}\right)_{e\text{-}e} + \left(\frac{d\mathcal{U}^{\mathrm{NT}}}{dt}\right)_{e\text{-ph}}, \quad \text{(B1b)}$$

as well as equations for the thermal electron energy,

$$C_e(T_e)\frac{dT_e}{dt} = \left(\frac{d\mathcal{U}^T}{dt}\right)_{e\text{-ph}} - \left(\frac{d\mathcal{U}^{\mathrm{NT}}}{dt}\right)_{e\text{-}e}, \quad \text{(B2)}$$

and phonon energy

$$C_{\mathrm{ph}}\frac{dT_{\mathrm{ph}}}{dt} = -\left(\frac{d\mathcal{U}^T}{dt}\right)_{e\text{-ph}} - \left(\frac{d\mathcal{U}^{\mathrm{NT}}}{dt}\right)_{e\text{-ph}}. \quad \text{(B3)}$$

C_e is the electron heat capacity, which can be determined by (see details in Ref. [12])

$$C_e(T_e) = \int \mathcal{E}\rho_e(\mathcal{E})\frac{\partial f^T(\mathcal{E}, \mu(T_e), T_e)}{\partial T_e}d\mathcal{E}, \quad \text{(B4)}$$

where $\mu(T_e)$ is the T_e-dependent chemical potential [12], which ensures the number conservation of the electron subsystem,

$$\int \rho_e(\mathcal{E})f^T(\mathcal{E}, \mu(T_e), T_e)d\mathcal{E} = \int \rho_e(\mathcal{E})f^T(\mathcal{E}, \mathcal{E}_F, 0\text{ K})d\mathcal{E}. \quad \text{(B5)}$$

The phonon heat capacity is taken to be $C_{\mathrm{ph}} = 2.54 \times 10^6$ J m^{-3} K^{-1} [57].

T_e in Eq. (B2) is referred to as the *instantaneous* electron temperature; it differs from the temperature extracted from the solution of BE (see Fig. 10 in Appendix C) because the instantaneous T_e has a longer rise time than the *extracted* T_e. $(\partial f/dt)_{\text{exc}}$ is the electron-photon excitation term given by Eq. (2), while $(\partial f^{\text{NT}}/\partial t)_{e\text{-}e}$ and $(\partial f^{\text{NT}}/\partial t)_{e\text{-ph}}$ are, respectively, the rates of nonthermal electron relaxation via the e-e and e-ph interactions. $p_{\text{abs}}(t)$ is the absorbed power density given by Eq. (6), $(d\mathcal{U}^{\text{NT}}/dt)_{e\text{-}e}$ is the nonthermal energy relaxation due to the e-e interaction; $(d\mathcal{U}^{\text{NT}}/dt)_{e\text{-ph}}$ and $\left(d\mathcal{U}^{T}/dt\right)_{e\text{-ph}}$ are, respectively, the energy transfer from the nonthermal and thermal electrons to phonons, and C_e and C_{ph} are, respectively, the heat capacity of the electron and phonon subsystems. In Eq. (B1a), we calculate the nonthermal part of the electron distribution explicitly since the nonthermal energy is not enough for calculating the permittivity, see Appendix C.

Unlike the rigorous approach of Eq. (1), to derive the ETTM, the e-e relaxation term is simplified using the relaxation time approximation, namely,

$$\left(\frac{\partial f^{\text{NT}}}{\partial t}\right)_{e\text{-}e} = -\frac{f^{\text{NT}}}{\tau_{e\text{-}e}}, \qquad \text{(B6)}$$

where $\tau_{e\text{-}e}$ is the e-e relaxation time. In many previous studies [4,9–11,58], $\tau_{e\text{-}e}$ was estimated using the Gurzhi formula [59–61],

$$\frac{1}{\tau_{e\text{-}e}(T_e)} = \frac{\omega_{\text{pump}}^2}{4\pi^2 \omega_p(T_e)}\left[1 + \left(\frac{2\pi k_B T_e}{\hbar \omega_{\text{pump}}}\right)^2\right], \qquad \text{(B7)}$$

where $\omega_p(T_e)$ is the plasma frequency given by Eq. (10) [2–10]. This formula is identical to the Fermi-liquid relaxation time with $\mathcal{E} = \mathcal{E}_F + \hbar\omega/2$; it is a good approximation for high-electron-density Drude materials such as noble metals. For LEDD materials with a non-parabolic conduction band such as ITO, the Gurzhi formula (B7) suggests that $\tau_{e\text{-}e}^{-1}$ exhibits a superquadratic growth with T_e (since the plasma frequency Eq. (10) decreases with T_e [3,4,9,11,31]). In contrast, in the ANTH$\mathscr{E}$M, we calculate the e-e relaxation time from the functional derivative of the e-e collision term. In particular, $\tau_{e\text{-}e}$ is shown [12] to be well described by the Fermi-liquid theory with a correction factor, i.e., $\tau_{e\text{-}e,\mathbf{k}}^{-1} \sim (1 + 2C\mu(T_e))^3\left[(\pi k_B T_e)^2 + (\mathcal{E}_{\mathbf{k}} - \mu(T_e))^2\right]$. Due to the decrease of the chemical potential with T_e [12] [see also Fig. 2(b)], it increases only nearly linearly with T_e (hence, much slower than that predicted by the Gurzhi formula). This finding justifies the phenomenological adjustment to the T_e dependence of $\tau_{e\text{-}e}^{-1}$ in Ref. [9].

$(\partial f^{\text{NT}}/\partial t)_{e\text{-ph}}$ and $(d\mathcal{U}^{\text{NT}}/dt)_{e\text{-ph}}$ are also treated using the relaxation time approximation in previous studies [4,10,11], i.e., $(\partial f^{\text{NT}}/\partial t)_{e\text{-ph}} = -f^{\text{NT}}/\tau_{e\text{-ph}}$ and

$(d\mathcal{U}^{\text{NT}}/dt)_{e\text{-ph}} = -\mathcal{U}^{\text{NT}}/\tau_{e\text{-ph}}$, where $\tau_{e\text{-ph}}$ is the e-ph relaxation time and is estimated using (e.g., Ref. [4])

$$\frac{1}{\tau_{e\text{-ph}}(T_{\text{ph}})} = \Gamma_0\left[\frac{2}{5} + \frac{4T_{\text{ph}}^5}{T_D^5}\int_0^{T_D/T_{\text{ph}}}\frac{z^4}{e^z - 1}dz\right], \quad \text{(B8)}$$

where $\Gamma_0 = 0.5342$ fs^{-1}, and $T_D = 900$ K is the Debye temperature. In the ANTH$\mathscr{E}$M, $\tau_{e\text{-ph}}$ is evaluated from the functional derivative of the e-ph collision term accounting for the momentum conservation, and is shown to be proportional to the phonon temperature and to be weakly dependent on T_e. The value of $\tau_{e\text{-ph}}$ evaluated using the explicit form given in Ref. [12] is similar to that obtained from Eq. (B8) and is shorter than 10 fs, i.e., much shorter than the e-e relaxation time. In this case, most of the nonthermal energy directly relaxes to the phonon subsystem after the generation of the nonthermal electron distribution, such that the electron subsystem will not heat up at all. This contradicts with the experimental observations, which typically conclude that the e-ph relaxation is much slower than the e-e relaxation (e.g., Ref. [9]). This disagreement can be reconciled by comparing the total energy transfer rate from electrons to phonons [denoted as $(d\mathcal{U}/dt)_{e\text{-ph}}$] with its contribution from the thermal electrons [i.e. $(d\mathcal{U}^{T}/dt)_{e\text{-ph}}$]. These quantities can be calculated from the e-ph collision term in Eq. (1) using the (total) electron distribution and a Fermi-Dirac distribution, namely,

$$\left(\frac{d\mathcal{U}}{dt}\right)_{e\text{-ph}} = \int \mathcal{E}\rho_e(\mathcal{E})\left(\frac{\partial f}{\partial t}\right)_{e\text{-ph}}d\mathcal{E} \qquad \text{(B9)}$$

and

$$\left(\frac{d\mathcal{U}^{T}}{dt}\right)_{e\text{-ph}} = \int \mathcal{E}\rho_e(\mathcal{E})\left(\frac{\partial f^{T}(\mathcal{E}, \mu(T_e), T_e)}{\partial t}\right)_{e\text{-ph}}d\mathcal{E}. \qquad \text{(B10)}$$

$(d\mathcal{U}^{\text{NT}}/dt)_{e\text{-ph}}$ is then the difference between $(d\mathcal{U}/dt)_{e\text{-ph}}$ and $(d\mathcal{U}^{T}/dt)_{e\text{-ph}}$, i.e.,

$$\left(\frac{d\mathcal{U}^{\text{NT}}}{dt}\right)_{e\text{-ph}} = \left(\frac{d\mathcal{U}}{dt}\right)_{e\text{-ph}} - \left(\frac{d\mathcal{U}^{T}}{dt}\right)_{e\text{-ph}}.$$

Our model shows that the total e-ph energy transfer rate is dominated by the thermal part of the electrons, i.e., $(d\mathcal{U}/dt)_{e\text{-ph}} \approx (d\mathcal{U}^{T}/dt)_{e\text{-ph}}$, see Fig. 9(b). This means that the contribution from the nonthermal part of the electrons $(d\mathcal{U}^{\text{NT}}/dt)_{e\text{-ph}}$ is, in fact, negligible (see also Refs. [9,15,25]), confirming the ad hoc neglect of $(d\mathcal{U}^{\text{NT}}/dt)_{e\text{-ph}}$ done in Ref. [9].

To have a deeper understanding of this, we cast the e-ph collision term into a RTA form that enforces number

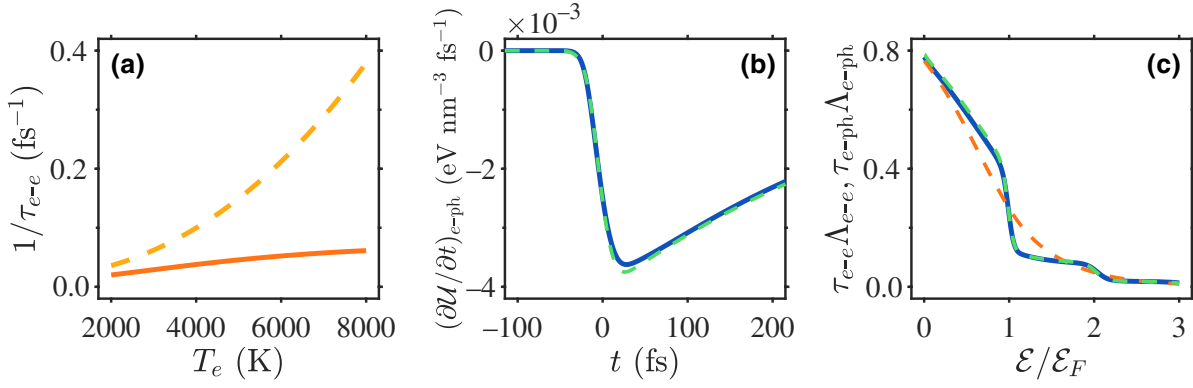FIG. 9. (a) The T_e dependence of the e-e relaxation rate $\tau_{e\text{-}e}^{-1}$. The orange solid line represents the result calculated using Eq. (17) in Ref. [12] at $\mathcal{E} = \mu(T_e) + \hbar\omega_{\text{pump}}/2$. The orange dashed line represents the result calculated using the Gurzhi formula (B7). (b) The rate of total energy transfer from the electron to the phonon subsystem Eq. (B9) (blue solid line). The green dashed line is the contribution from the thermal part of the electron distribution (B10). The results correspond to the case of $I_0 = 161$ GW/cm^2 and $\tau_{\text{pump}} = 30$ fs. (c) The electron energy dependence of $\Lambda_{e\text{-ph}}\tau_{e\text{-ph}}$ [see Eq. (B11)] (green dashed line) and $\Lambda_{e\text{-}e}\tau_{e\text{-}e}$ (orange dashed line) for the same case as (b) at $t = 0$ fs. The blue solid line is the nonequilibrium electron distribution in Eq. (B11) at $t = 0$ fs.

conservation (by adding the Lorentz term, see Ref. [25]), namely,

$$(\partial f/\partial t)_{e\text{-ph}} = -f/\tau_{e\text{-ph}} + \Lambda_{e\text{-ph}}. \quad \text{(B11)}$$

Thus, $\tau_{e\text{-ph}}\Lambda_{e\text{-ph}}$ can be interpreted as the target distribution, which the nonequilibrium electron distribution is going to relax to via the e-ph interaction if the excitation is stopped and the e-e interaction is turned off. Figure 9(c) shows that the target distribution has a nonthermal shoulder structure and differs from the distribution before relaxation only slightly, thus is distinct from any thermal distribution. In contrast, if we perform the same analysis for the e-e collision term, we find that the target distribution is similar to a thermal distribution with some high temperature, see Fig. 9(c). Moreover, the e-e Lorentz term is 2 orders of magnitude larger than the e-ph Lorentz term. This explains why the e-e relaxation is much faster than the e-ph relaxation [$\left|(\partial f/\partial t)_{e\text{-}e}\right| \gg \left|(\partial f/\partial t)_{e\text{-ph}}\right|$, see Fig. 2 in Ref. [12]] although $\tau_{e\text{-}e}^{-1} < \tau_{e\text{-ph}}^{-1}$. The comparison above further indicates that the e-ph interaction is mainly responsible for transferring the thermal energy of the electrons to phonon, instead of thermalizing the nonequilibrium electron distribution. Therefore, the e-ph interaction violates the RTA assumption, so that it is improper to apply the relaxation time approximation for the e-ph relaxation of the nonthermal electron energy. This conclusion is also valid for noble metals.

The energy transfer rate from the thermal electrons to the phonons $(d\mathcal{U}^T/dt)_{e\text{-ph}}$ [Eq. (B10)] has been shown in Ref. [12] to be proportional to the temperature difference between electrons and phonons, namely, $(d\mathcal{U}^T/dt)_{e\text{-ph}} = -G_{e\text{-ph}}(T_e - T_{\text{ph}})$, where $G_{e\text{-ph}}$ is the electron-phonon energy coupling coefficient. The electron-phonon energy coupling coefficient was previously evaluated using the

formulation derived for noble metals (e.g., in Refs. [4,9, 11]). However, we showed in Ref. [12] that this approach overestimates the correct value by more than an order of magnitude due to the failure to account for momentum conservation in the e-ph interaction. This explains the need to reduce the e-ph coupling coefficient by a factor of 8 in Ref. [9] to fit the experimental data.

Finally, if the e-e relaxation time is approximated to be zero, i.e., the electron subsystem is assumed to become thermalized instantaneously, Eq. (B1) reduces to the two-temperature model [62],

$$C_e(T_e)\frac{dT_e}{dt} = -G_{e\text{-ph}}(T_e - T_{\text{ph}}) + P_{\text{abs}}(t), \quad \text{(B12a)}$$

$$C_{\text{ph}}\frac{dT_{\text{ph}}}{dt} = G_{e\text{-ph}}(T_e - T_{\text{ph}}). \quad \text{(B12b)}$$

Here, T_e is referred to as the *effective* electron temperature; it differs from the extracted and *instantaneous* T_e since thermalization is assumed to be instantaneous in Eq. (B12).

As shown in this work and others [9,12], the TTM can be quite accurate for pulses having duration longer than the e-e relaxation time.

APPENDIX C: LIMITATION OF USING THERMAL DISTRIBUTION TO EVALUATE THE PERMITTIVITY

In this section, we use the example of the 30-fs pump pulse with $I_0 = 161$ GW/cm^2 studied in Sec. III E to demonstrate the relevance of accounting for the nonthermal part of the electron distribution in the permittivity calculation when self-consistently solving the ETTM [Eqs. (B1)–(B3)] coupled with Maxwell's equation (4).
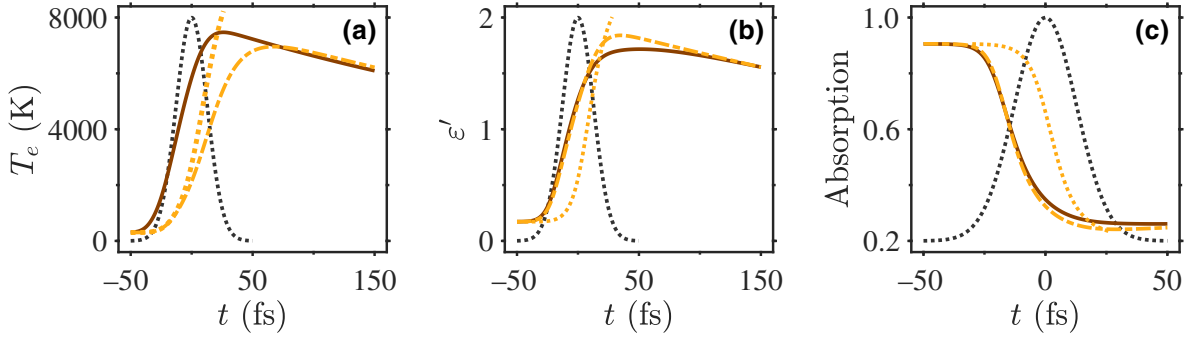
FIG. 10. (a),(b) the same as Figs. 8(b) and 8(d) for the 30-fs pulse with $I_0 = 161$ GW/cm^2. (c) The absorption as a function of time for the same case as (a),(b). The dash-dotted and dotted lines represent the results obtained from ETTM with the nonthermal and thermal permittivity model, respectively. The black dotted lines represent the temporal intensity profile of the incident pump pulse.

To do that, we evaluate the permittivity using two different methods in the self-consistent calculation. In the first case, the ITO permittivity [Eq. (5)] is evaluated using the electron distribution $f(\mathcal{E}) = f^T(\mathcal{E}, \mu(T_e), T_e) + f^{NT}(\mathcal{E})$. In the second one, we neglect $f^{NT}(\mathcal{E})$ and evaluate the ITO permittivity using only the thermal distribution $f(\mathcal{E}, \mu(T_e), T_e)$. For the first case, the instantaneous T_e dynamics is qualitatively similar to that of the extracted T_e except that at the early stages, the rise of the instantaneous T_e is controlled by e-e relaxation time, which is slower than the pulse duration, see Fig. 10(a). In contrast, the rise time of ε' (and thus the decrease of absorption) are controlled by the pulse duration, showing remarkable agreement with the results obtained from our ANTH$\mathscr{E}$M [see Figs. 10(b) and 10(c)].

However, if one neglects the nonthermal part of the electron distribution and simply evaluates the ITO permittivity using a thermal distribution with instantaneous T_e, the rise time of the permittivity (and the decrease of absorption) will be the same as instantaneous T_e thus will be controlled by the e-e relaxation time instead of the pulse duration, see the dotted lines in Fig. 10. This causes the maximum value of the instantaneous T_e to be $> 10^4$ K, much higher than the result of the case accounting for the nonthermal distribution in the permittivity calculation. This comparison clearly shows that the nonthermal electron distribution is essential to properly capture the nonlinear response of ITO.

The need to account for the nonthermal distribution also occurred when calculating the change of the contribution of interband transitions to the permittivity of noble metals under ultrafast laser pulse excitation. In that case, the nonthermal electron distribution is assumed to have the same profile as the photon excitation term, i.e., $f^{NT}(\mathcal{E}, t) = A(t)[f^T(\mathcal{E} - \hbar\omega_{\text{pump}})(1 - f^T(\mathcal{E})) - f^T(\mathcal{E})(1 - f^T(\mathcal{E} + \hbar\omega_{\text{pump}}))]$, where $A(t)$ is determined from the nonthermal energy $\mathcal{U}^{NT}(t)$ [38,58,63]. This is a decent approximation for noble metals because the chemical potential is nearly T_e

independent. Unfortunately, this approximation is not valid for ITO since the chemical potential decreases significantly with T_e, thus, requiring the use of the full nonthermal model (ANTH$\mathscr{E}$M).

APPENDIX D: PROBE PULSE DYNAMICS FOR THE SHORTER PUMP PULSES

In this Appendix, we show the comparison of the reflection and spectrum of the reflected probe pulse (with duration of 220 fs) among the three cases studied in Sec. III E. Figure 11(a) shows that decay rate of the probe-pulse reflection for the case of 30-fs pump pulse with $I_0 = 161$ GW/cm^2 is (almost) the same as the case of 220-fs
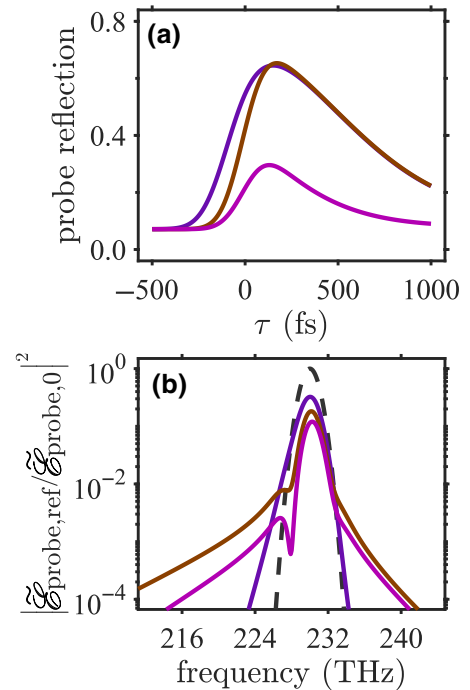

FIG. 11. The same as Figs. 7(a) and 7(c) but for cases of shorter pump pulses studied in Fig. 8 (the same color is used).

pump pulse with $I_0 = 22$ GW/cm^2, and is slower than that the case of 30-fs pump pulse with $I_0 = 22$ GW/cm^2. This indicates that the decay rate of the probe-pulse reflection is slower for stronger pump-pulse energy but is weakly sensitive to the pump-pulse duration. In contrast, the spectrum of the reflected probe pulse is wider for shorter pump-pulse duration, see Fig. 11(b).

---

[1] N. Kinsey and J. Khurgin, Nonlinear epsilon-near-zero materials explained: Opinion, Opt. Mater. Express **9,** 2793 (2019).

[2] Q. Guo, Y. Cui, Y. Yao, Y. Ye, Y. Yang, X. Liu, S. Zhang, X. Liu, J. Qiu, and H. Hosono, A solution-processed ultrafast optical switch based on a nanostructured epsilon-near-zero medium, Adv. Mater. **29,** 1700754 (2017).

[3] P. Guo, R. D. Schaller, J. B. Ketterson, and R. P. H. Chang, Ultrafast switching of tunable infrared plasmons in indium tin oxide nanorod arrays with large absolute amplitude, Nat. Photon. **10,** 267 (2016).

[4] M. Z. Alam, S. A. Schulz, J. Upham, I. D. Leon, and R. W. Boyd, Large optical nonlinearity of nanoantennas coupled to an epsilon-near-zero material, Nat. Photon. **12,** 79 (2018).

[5] R. Secondo, J. Khurgin, and N. Kinsey, Absorptive loss and band non-parabolicity as a physical origin of large nonlinearity in epsilon-near-zero materials, Opt. Mater. Express **10,** 1545 (2020).

[6] H. Wang, K. Du, R. Liu, X. Dai, W. Zhang, S. J. Chua, and T. Mei, Role of hot electron scattering in epsilon-near-zero optical nonlinearity, Nanophotonics **9,** 4287 (2020).

[7] J. Bohn, T. S. Luk, C. Tollerton, S. Hutchins, I. Brener, S. Horsley, W. L. Barnes, and E. Hendry, All-optical switching of an epsilon-near-zero plasmon resonance in indium tin oxide, Nat. Commun. **12,** 1017 (2021).

[8] R. Tirole, E. Galiffi, J. Dranczewski, T. Attavar, B. Tilmann, Y.-T. Wang, P. A. Huidobro, A. Alú, J. B. Pendry, S. A. Maier, S. Vezzoli, and R. Sapienza, Saturable Time-Varying Mirror Based on an Epsilon-Near-Zero Material, Phys. Rev. Appl. **18,** 054067 (2022).

[9] J. Baxter, A. Pérez-Casanova, L. Cortes-Herrera, A. Calá Lesina, I. De Leon, and L. Ramunno, Dynamic nanophotonics in epsilon-near-zero conductive oxide films and metasurfaces: A quantitative, nonlinear, computational model, Adv. Photon. Res. **n/a,** 2200280 (2023).

[10] M. Z. Alam, I. D. Leon, and R. W. Boyd, Large optical non-linearity of indium tin oxide in its epsilon-near-zero region, Science **116,** 795 (2016).

[11] E. Minerbi, S. Sideris, J. B. Khurgin, and T. Ellenbogen, The role of epsilon near zero and hot electrons in enhanced dynamic THz emission from nonlinear metasurfaces, Nano Lett. **22,** 6194 (2022). pMID: 35899937,.

[12] S. Sarkar, I. W. Un, and Y. Sivan, The Electronic and Thermal Response of Low Electron Density Drude Materials to Ultrafast Optical Illumination, Phys. Rev. Appl. **19,** 014005 (2022).

[13] In particular, the dependence of the *e-e* collision rate on the energy difference of the colliding electrons is fixed to

[14] H. Wang, K. Du, C. Jiang, Z. Yang, L. Ren, W. Zhang, S. J. Chua, and T. Mei, Extended Drude Model for Intraband-Transition-Induced Optical Nonlinearity, Phys. Rev. Appl. **11,** 064062 (2019).

[15] Y. Sivan, I. W. Un, and Y. Dubi, Assistance of plasmonic nanostructures to photocatalysis - just a regular heat source, Faraday Discuss. **214,** 215 (2019).

[16] R. W. Boyd, *Nonlinear Optics* (Academic Press, San Diego, 2003), 2nd ed.

[17] J. Gordon and R. Ziolkowsky, The design and simulated performance of a coated nano-particle laser, Opt. Exp. **15,** 2622 (2007).

[18] Y. Sivan, S. Xiao, U. K. Chettiar, A. V. Kildishev, and V. M. Shalaev, Frequency-domain simulations of a negative-index material with embedded gain, Opt. Exp. **17,** 24060 (2009).

[19] J. B. Khurgin, M. Clerici, and N. Kinsey, Fast and slow nonlinearities in epsilon-near-zero materials, Laser Photon. Rev. **15,** 2000291 (2021).

[20] T. Stoll, P. Maioli, A. Crut, N. D. Fatti, and F. Vallée, Advances in femto-nano-optics: Ultrafast nonlinearity of metal nanoparticles, Eur. Phys. J. B **87,** 260 (2014).

[21] I. Gurwich and Y. Sivan, A metal nanosphere under intense continuous wave illumination - a unique case of non-perturbative nonlinear nanophotonics, Phys. Rev. E **96,** 012212 (2017).

[22] I. W. Un and Y. Sivan, The thermo-optic nonlinearity of single metal nanoparticles under intense continuous-wave illumination, Phys. Rev. Mater. **4,** 105201 (2020).

[23] N. D. Fatti, C. Voisin, M. Achermann, S. Tzortzakis, D. Christofilos, and F. Valleé, Nonequilibrium electron dynamics in noble metals, Phys. Rev. B **61,** 16956 (2000).

[24] M. Kornbluth, A. Nitzan, and T. Seidman, Light-induced electronic non-equilibrium in plasmonic particles, J. Chem. Phys. **138,** 174707 (2013).

[25] Y. Dubi and Y. Sivan, "Hot electrons" in metallic nanostructures - non-thermal carriers or heating?, Light: Sci. Appl. **8,** 89 (2019).

[26] S. Sarkar, I. W. Un, Y. Sivan, and Y. Dubi, Theory of non-equilibrium "hot" carriers in direct band-gap semiconductors under continuous illumination, New J. Phys. **24,** 053008 (2022).

[27] For the example analyzed in this work where the carrier frequency is 230 THz, a time resolution of < 0.2 fs would be required.

[28] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Harcourt, Inc., New York, 1976).

[29] E. O. Kane, Band structure of indium antimonide, J. Phys. Chem. Solids **1,** 249 (1957).

[30] X. Liu, J. Park, J.-H. Kang, H. Yuan, Y. Cui, H. Y. Hwang, and M. L. Brongersma, Quantification and impact of non-parabolicity of the conduction band of indium tin oxide on its plasmonic properties, Appl. Phys. Lett. **105,** 181117 (2014).

[31] P. Guo, R. D. Schaller, L. E. Ocola, B. T. Diroll, J. B. Ketterson, and R. P. H. Chang, Large optical nonlinearity of ITO nanorods for sub-picosecond all-optical modulation of the full-visible spectrum, Nat. Commun. **7,** 12892 (2016).

[32] Note that within our formulation we can take any envelope as long as the pulse duration is much longer than $1/\eta$ in Eq. (9).

[33] T. G. Mackay and A. Lakhtakia, *The Transfer-Matrix Method in Electromagnetics and Optics* (Morgan & Claypool, San Rafael, CA, 2020).

[34] M. Born, E. Wolf, A. B. Bhatia, P. C. Clemmow, D. Gabor, A. R. Stokes, A. M. Taylor, P. A. Wayman, and W. L. Wilcock, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Cambridge University Press, Cambridge, 1999), 7th ed.

[35] The electron temperature decay rate can be obtained by setting $p_{abs} = 0$ in Eq. (B12a). Since the change of T_{ph} is much smaller than that of T_e, we set $T_{ph} = 300$ K when calculating the electron-temperature decay rate in the inset of Fig. 2(a).

[36] L. Caspani, R. P. M. Kaipurath, M. Clerici, M. Ferrera, T. Roger, J. Kim, N. Kinsey, M. Pietrzyk, A. D. Falco, V. M. Shalaev, A. Boltasseva, and D. Faccio, Enhanced Nonlinear Refractive Index in ϵ-Near-Zero Materials, [Phys. Rev. Lett. **116**, 233901 (2016).](#)

[37] Y. Yang, K. Kelley, E. Sachet, S. Campione, T. S. Luk, J.-P. Maria, M. B. Sinclair, and I. Brener, Femtosecond optical polarization switching using a cadmium oxide-based perfect absorber, [Nat. Photon. **11**, 390 (2017).](#)

[38] M. Zavelani-Rossi, D. Polli, S. Kochtcheev, A.-L. Baudrion, J. Béal, V. Kumar, E. Molotokaite, M. Marangoni, S. Longhi, G. Cerullo, P.-M. Adam, and G. Della Valle, Transient optical response of a single gold nanoantenna: The role of plasmon detuning, [ACS Photon. **2**, 521 (2015).](#)

[39] Here, $\alpha \sim - \left. \partial \omega_p^2 / \partial T_e \right|_{T_e = T_0} > 0$ represents the decrease of the effective plasma frequency with T_e.

[40] This is because electrons stop transferring their energy once T_e and T_{ph} are the same. At this stage, T_e is higher than its initial temperature T_0. This means that a small amount of the absorbed energy remains in the electron subsystem (this amount of the energy is small since $C_e \ll C_{ph}$). This excess energy will be transferred to the environment on the slower timescale neglected in this work.

[41] Therefore, one should not compare Figs. 6(a) with 3 in [8].

[42] R. H. M. Groeneveld, R. Sprik, and A. Lagendijk, Femtosecond spectroscopy of electron-electron and electron-phonon energy relaxation in Ag and Au, [Phys. Rev. B **51**, 11433 (1995).](#)

[43] Y. Ishida, T. Togashi, K. Yamamoto, M. Tanaka, T. Taniuchi, T. Kiss, M. Nakajima, T. Suemoto, and S. Shin, Non-thermal hot electrons ultrafastly generating hot optical phonons in graphite, [Sci. Rep. **1**, 64 (2011).](#)

[44] R. Tirole, S. Vezzoli, E. Galiffi, I. Robertson, D. Maurice, B. Tilmann, S. A. Maier, J. B. Pendry, and R. Sapienza, Double-slit time diffraction at optical frequencies, (2022), [ArXiv:2206.04362.](#)

[45] M. Massicotte, G. Soavi, A. Principi, and K.-J. Tielrooij, Hot carriers in graphene–fundamentals and applications, [Nanoscale **13**, 8376 (2021).](#)

[46] M. Z. Hasan and J. E. Moore, Three-dimensional topological insulators, [Annu. Rev. Condens. Matter Phys. **2**, 55 (2011).](#)

[47] B. T. Diroll, S. Saha, V. M. Shalaev, A. Boltasseva, and R. D. Schaller, Broadband ultrafast dynamics of refractory metals: TiN and ZrN, [Adv. Opt. Mater. **8**, 2000652 (2020).](#)

[48] G. V. Naik, J. Kim, and A. Boltasseva, Oxides and nitrides as alternative plasmonic materials in the optical range, [Opt. Mater. Express **1**, 1090 (2011).](#)

[49] J. Wu, Z. T. Xie, Y. Sha, H. Y. Fu, and Q. Li, Epsilon-near-zero photonics: Infinite potentials, [Photon. Res. **9**, 1616 (2021).](#)

[50] O. Reshef, E. Giese, M. Z. Alam, I. D. Leon, J. Upham, and R. W. Boyd, Beyond the perturbative description of the nonlinear optical response of low-index materials, [Opt. Lett. **42**, 3225 (2017).](#)

[51] A. Marini, M. Conforti, G. D. Valle, H. W. Lee, T. X. Tran, W. Chang, M. A. Schmidt, S. Longhi, P. S. J. Russell, and F. Biancalana, Ultrafast nonlinear dynamics of surface plasmon polaritons in gold nanowires due to the intrinsic nonlinearity of metals, [New J. Phys. **15**, 013033 (2013).](#)

[52] R. W. Boyd, Z. Shi, and I. D. Leon, The third-order nonlinear optical susceptibility of gold, [Opt. Comm. **326**, 74 (2014).](#)

[53] Y. Zhou, M. Z. Alam, M. Karimi, J. Upham, O. Reshef, C. Liu, A. E. Willner, and R. W. Boyd, Broadband frequency translation through time refraction in an epsilon-near-zero material, [Nat. Commun. **11**, 2180 (2020).](#)

[54] S. J. Orfanidis, Electromagnetic waves and antennas (2002).

[55] L. D. Landau, J. Bell, M. Kearsley, L. Pitaevskii, E. Lifshitz, and J. Sykes, *Electrodynamics of Continuous Media* Vol. 8 (Pergamon Press Ltd., Oxford, England, 1984), 2nd ed.

[56] For ITO, the damping rate is around 0.1 fs^{-1}, see Fig. 1 in Ref. [12]. From Eqs. (2)–(3), it follows that the condition $\left| (df/dt)_{exc} \right| < \eta$ is therefore satisfied when the local field is smaller than approximately 2.7×10^9 V/m, corresponding to an incident intensity of approximately 200 GW/cm^2 for the example analyzed in this work. Since the absorption decreases as T_e increases, a much higher incident intensity than approximately 200 GW/cm^2 is required to violate $\left| (df/dt)_{exc} \right| < \eta$ at higher T_e.

[57] T. Yagi, K. Tamano, Y. Sato, N. Taketoshi, T. Baba, and Y. Shigesato, Analysis on thermal properties of tin doped indium oxide films by picosecond thermoreflectance measurement, [J. Vac. Sci. Technol. A **23**, 1180 (2005).](#)

[58] E. Carpene, Ultrafast laser irradiation of metals: Beyond the two-temperature model, [Phys. Rev. B **74**, 024301 (2006).](#)

[59] R. N. Gurzhi, Mutual electron correlations in metal optics, Sov. Phys. JETP **105**, 673 (1959).

[60] J. B. Smith and H. Ehrenreich, Frequency dependence of the optical relaxation time in metals, [Phys. Rev. B **25**, 923 (1982).](#)

[61] C. Voisin, N. Del Fatti, D. Christofilos, and F. Vallée, Ultrafast electron dynamics and optical nonlinearities in metal nanoparticles, [J. Phys. Chem. B **105**, 2264 (2001).](#)

[62] S. I. Anisimov, B. L. Kapeilovich, and T. I. Perelman, Electron emission from metal surfaces exposed to ultrashort laser pulses, Sov. Phys. JETP **39**, 375 (1974).

[63] C. K. Sun, F. Valleé, L. H. Acioli, E. P. Ippen, and J. G. Fujimoto, Femtosecond tunable measurement of electron thermalization in gold, [Phys. Rev. B **50**, 15337 (1994).](#)