An electronic-based model of the optical nonlinearity of low 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
(Dated: November 4, 2022)

Abstract

Low electron density Drude (LEDD) materials such as indium tin oxide (ITO) are receiving considerable attention because of their combination of CMOS compatibility, unique epsilon-near-zero (ENZ) behavior, and giant ultrafast nonlinear thermo-optic response. However, the understanding of the electronic and optical response of LEDD materials is so far based on simplistic extensions of known models of noble metals, frequently without the inclusion of the interplay among the lower electron density, relatively high Debye energy, and the non-parabolic band structure. To bridge this knowledge gap, this work provides 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 the newly derived time-dependent permittivity model for LEDD materials under optical pumping within the adiabatic approximation. We find that unlike noble metals, the electron temperatures can reach the Fermi temperature, in which case the effective chemical potential dramatically decreases and even becomes negative, thus, transiently converting the Drude metal into 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 because of changes to population. This resolves the argument about the rise time of the permittivity and shows that it is instantaneous. In this vein, we show that referring to the LEDD permittivity as having a “saturable” nonlinearity is unsuitable since its permittivity dynamics 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.

* iengwai@post.bgu.ac.il; These two authors have contributed equally.
I. INTRODUCTION

The technologically-important transparent conducting oxide ITO has recently emerged as an promising candidate for high efficiency nonlinear optics applications. The main reason is the unique epsilon-near-zero (ENZ) point it has in the near IR [1], which is characteristic of Low Electron Density Drude (LEDD) materials.

The nonlinear sub-picosecond optical response of LEDD materials was described using a variety of approaches, starting from a phenomenological temporally delayed response. Specifically, most studies described the LEDD material permittivity as having a time-dependent effective mass (and hence, plasma frequency), e.g., [2–10]. This is an effective description of the transient occupation of high energy electron states for which the effective mass is growing.

More recent models employed a Relaxation Time Approximation (RTA)-based extended Two Temperature Model (eTTM) [4, 5, 10, 11]. However, in the latter, some of the eTTM parameters were computed from equations derived for parabolic bands, the importance of momentum conservation to the $e-ph$ interactions was not yet known [12], the $e-e$ interaction were accounted for in a simplistic manner [13], it was not shown how the phonon temperature dynamics affects the ITO permittivity, and the analysis of the nonlinear response was specific for the experimental configuration. Thus, for example, the computed eTTM parameters in [10] (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 firm theoretical support for some of the values of parameters in use.

Notably, there are two additional limitations of the state-of-the-art nonlinear optical theory of LEDD materials. First, an important aspect of the dynamics which is not yet well-understood is the response to high illumination intensities. In [3, 9, 14], this response was referred to as saturable; the saturation intensity was claimed to be $\sim 100 \text{ GW/cm}^2$ [9], 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 gradual depletion of the (electronic) ground state of the material, an effect which leads to a decrease of the imaginary part of its permittivity (saturable absorption) [16], and even to the a change of its sign (upon pop-
ulation inversion) [17, 18]. This behaviour is in contradiction to the experimental results that the imaginary part of ITO permittivity grows upon illumination, see [10, 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 [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 importance 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, the simulations in [12] were limited to modestly high intensities.

In this work, we go beyond [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 Section II, we recall briefly the electronic model developed in [12], and describe the time-dependent permittivity model used in this work. In particular, we model the permittivity using an extension of the Lindhard formula, and describe a self-consistent approach for the calculation of the distribution, permittivity and local field. To our knowledge, this provides the first ever extension of the Lindhard formula to materials with a non-parabolic conduction band under high-intensity pulsed illumination.

As a specific example, we then focus on a prototypical geometry of an ITO layer illuminated by an obliquely incident pump pulse. In Section III A, we discuss the resulting dynamics of the electron and phonon temperatures as a function of illumination intensity. We show that the electron system can reach the Fermi temperature and the chemical potential can become negative. We also show that the decay rate of the electron temperature becomes faster with the illumination intensity.

In Section 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 (spectrally shifting the
resonance away from the ENZ point) and that it is dominated by changes to the electron temperature. 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 reported in [9]. Most importantly, these results show that the ITO permittivity dynamics is characteristic of a high quality metal (as discussed in [19] for pulsed illumination, and even in [20, 21] 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 behaviour. 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 thermalization time. We also explain the physical origin of the ad hoc corrections needed previously. Then, in Section 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 Section 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 peak intensity. This result reaches quantitative agreement with the experimental observation in [9]. We also study the nonlinear optical response to shorter pulses in Section III E and show that one needs to go beyond the TTM when the pulse duration is shorter than the thermalization time, by accounting for the non-thermal part of the distribution. Finally, we provide a discussion of the results and an outlook in Section IV.

II. FORMULATION

A. Model for the electron distribution dynamics and phonon temperature

We start by revisiting and extending the theoretical approach presented in [12]. In particular, we solve the Boltzmann equation (BE) for the electron dynamics, namely,
\[ \frac{\partial f(\mathcal{E}, t)}{\partial t} = \left( \frac{\partial f(\mathcal{E}, t; \varepsilon(t, \omega_{\text{pump}}), |\mathbf{E}(t, \omega_{\text{pump}})|^2)}{\partial t} \right)_{\text{exc}} + \left( \frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{e-\text{ph collision}} + \left( \frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{e-e\ collision} + \left( \frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{e-\text{imp collision}}. \]  

(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 (\( e-\text{ph} \)) and electron-charged-impurity (\( e-\text{imp} \)) interactions are the same as in [12, 22–24], the excitation term is more complicated. In particular, the electron population evolves in time such that it is given by

\[ \left( \frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{\text{exc}} = B(t; \omega_{\text{pump}}) \left[ D_J(\mathcal{E}-\hbar\omega_{\text{pump}}, \mathcal{E}) \rho_e(\mathcal{E}-\hbar\omega_{\text{pump}}) f(\mathcal{E}-\hbar\omega_{\text{pump}}, t)(1-f(\mathcal{E}, t)) 
- D_J(\mathcal{E}, \mathcal{E}+\hbar\omega_{\text{pump}}) \rho_e(\mathcal{E}+\hbar\omega_{\text{pump}}) f(\mathcal{E}, t)(1-f(\mathcal{E}+\hbar\omega_{\text{pump}}, t)) \right]. \]  

(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 \( \mathcal{E}_{\text{initial}} \rightarrow \mathcal{E}_{\text{final}} \). 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 \( P_{\text{abs}} \), namely,

\[ \int \mathcal{E} \rho_e(\mathcal{E}) \left( \frac{\partial f(\mathcal{E}, t)}{\partial t} \right)_{\text{exc}} d\mathcal{E} = P_{\text{abs}}(t; \omega_{\text{pump}}). \]  

(3)

Here, the absorbed power density is evaluated dynamically and consistently either \( f, \varepsilon \) and \( \mathbf{E} \) using the Poynting’s theorem (see details in Section 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 [24, 25]), 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 on 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 calculation can be very time consuming and computationally expensive [26]. 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 $\sim 2.7 \times 10^9$ V/m. For this purpose, we further assume that the pump pulse duration is much longer than the periodicity so that the electric field $E(t)$ (and the electric displacement $D(t)$) can be considered quasi-monochromatic and be written as a product of a slowly varying envelope ($E(t)$ or $D(t)$) and a rapidly varying phase factor, namely, $E(t) = E(t)e^{-i\omega_{\text{pump}}t} + c.c.$ ($D(t) = D(t)e^{-i\omega_{\text{pump}}t} + c.c.$). We show in Appendix A that the electric displacement envelope is then related to the local field envelope by $D(t) = \varepsilon_0\varepsilon(t;\omega_{\text{pump}})E(t)$, and that Ampère’s law and Maxwell–Faraday equation become

$$
\nabla \times H(r, t) = -i\omega_{\text{pump}}\varepsilon_0\varepsilon(t;\omega_{\text{pump}})E(r, t)
$$

$$
\nabla \times E(r, t) = -i\omega_{\text{pump}}\mu_0 H(r, t)
$$

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

$$
\varepsilon(t;\omega_{\text{pump}}) = \varepsilon_\infty + \lim_{q \to 0} \frac{2e^2}{\varepsilon_0 q^2} \int \frac{d^3k}{(2\pi)^3} \frac{f_{k+q}(t;\omega_{\text{pump}}) - f_k(t;\omega_{\text{pump}})}{\varepsilon_k + q - i\hbar(\eta_{k+q}(t;\omega_{\text{pump}}) + \eta_k(t;\omega_{\text{pump}}))}.
$$

where $\varepsilon_\infty$ represents the contribution of interband transitions to the permittivity, $f_k$ is the momentum ($k$)-dependent electron distribution in the conduction band, $\varepsilon_k$ is the electron energy satisfying the $\varepsilon - k$ relation (see Eq. [7] below), $q$ is the wavevector of the applied field (taken to zero limit because it is much smaller than the electron wavevector), and

$$
\eta_k = \tau_{e-e,k}^{-1} + \tau_{e-ph,k}^{-1} + \tau_{e-imp,k}^{-1}
$$

is the damping rate following from Matthiessen’s rule [27]. $\tau_{e-e,k}^{-1}$, $\tau_{e-ph,k}^{-1}$ and $\tau_{e-imp,k}^{-1}$ are the collision rates associated with the $e-e$, $e-ph$ and $e-imp$ interactions, respectively (see details in [12]). In addition, the contribution of the interband transition ($\varepsilon_\infty$) is assumed to be independent of time, temperature and electric field and not to contribute to the dispersion at frequency below the interband threshold of $\sim 3.2$ eV.
The local field envelope (the solution of Eqs. (4)) and the time-dependent permittivity (5) are next used to calculate the power absorbed density via the Poynting theorem

\[ P_{\text{abs}}(t; \omega_{\text{pump}}) = \frac{\omega_{\text{pump}}^2}{2} \varepsilon_0 \varepsilon''(t; \omega_{\text{pump}}) |E(t; \omega_{\text{pump}})|^2. \]  

(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 Non-Thermal permittivity (\( \varepsilon \)) Model (ANTH\( \varepsilon \)M).

It is then customary to rewrite the integral in energy space. For materials with a non-parabolic conduction band, e.g., in ITO, the energy-momentum (\( \mathcal{E} - k \)) relation is expressed by the Kane quasi-linear dispersion [28],

\[ \hbar^2 k^2 = 2m_e^* \mathcal{E}_k (1 + C \mathcal{E}_k), \]  

(7)

such that the electron density of states (eDOS) is given by [12, 28, 29]

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

(8)

where \( m_e^* = 0.3964m_e \) is the electron effective mass at the conduction band minimum, and \( C = 0.4191 \text{ eV}^{-1} \) [29] is the first-order non-parabolicity 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

\[ \varepsilon(t; \omega_{\text{pump}}) = \varepsilon_\infty - \frac{e^2}{\varepsilon_0 m_e^*} \int \frac{\rho_e(\mathcal{E}) f(\mathcal{E}, t; \omega_{\text{pump}})}{(\omega_{\text{pump}} + i\eta(\mathcal{E}, t; \omega_{\text{pump}}))^2} \left( 1 + \frac{8C\mathcal{E}(1 + C\mathcal{E})/3}{(1 + 4C\mathcal{E}(1 + C\mathcal{E}))^{3/2}} \right) d\mathcal{E}. \]  

(9)

Eq. (9) represents a rigorous extension of the Lindhard’s formula to time-varying systems having a non-parabolic conduction band; this account is first of its kind in the context of ITO modelling. In particular, the permittivity [9] includes the contribution from the non-thermal 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 \( \tau_{e-e} \) such as those used in [2-6, 8-10], but that it is non-negligible for pulse durations comparable or shorter than \( \tau_{e-e} \). Moreover, the damping rate \( \eta \) in Eq. (9) is obtained self-consistently from the electron distribution via the \( e - e \), \( e - ph \) and \( e - imp \) scattering rates (see [12]). This is in contrast with the phenomenological methods
addressing the permittivity damping term in previous studies \[3, 5, 10, 11, 30\]. Below, we show that this self-consistent form successfully explains the experimental data \[4, 9, 14\] without introducing phenomenological adjustments.

If the damping term $\eta$ is approximated to be energy-independent, then, Eq. (9) reproduces the Drude formula for the relative permittivity with an electron-distribution-dependent plasma frequency \[3, 5, 10, 11, 30, 31\], namely,

$$\omega_p^2[f(E)] = \frac{e^2}{3\pi^2\varepsilon_0 m_e^*} \int \left[ \frac{2m_e^* E}{\hbar^2} (1 + CE) \right]^{3/2} (1 + 2CE)^{-1} \left( -\frac{\partial f}{\partial E} \right) dE. \quad (10)$$

The extra factor for $C \neq 0$ in the integral of Eq. (9) 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 non-parabolicity.

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. 3(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. 4); 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., \[20, 21\]), 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 [9], where a pump and a probe pulse are obliquely incident on a sample consisting of a 40 nm thin film of ITO deposited on glass and covered by a 100 nm gold film, see Fig. 1. This configuration is effective mainly because of the low wave impedance of the incoming pulse which enables good light penetration into the ITO.

FIG. 1. (Color online) Schematic diagram of the setup considered here and in experiment [9]. 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 were obliquely incident on the sample at angles of 60° and 65°, respectively. The probe pulse arrives at the sample with a time delay $\tau$ after the pump pulse.

The electric field of the incident pump and probe pulsed illumination are given as the product of an envelope and a carrier wave, $E_{\text{pump}}(t) = (E_{\text{pump},0}/2)e^{-2\ln 2(t/\tau_{\text{pump}})^2}e^{-i\omega_{\text{pump}}t} + c.c.$ and $E_{\text{probe}}(t; \tau) = (E_{\text{probe},0}/2)e^{-2\ln 2((t-\tau)/\tau_{\text{probe}})^2}e^{-i\omega_{\text{probe}}(t-\tau)} + c.c.$, where $E_{\text{pump},0}$ is related to the peak intensity by $I_0 = \frac{1}{2}\sqrt{\frac{\varepsilon_0}{\mu_0}}|E_{\text{pump},0}|^2$; $\tau_{\text{pump}} = \tau_{\text{probe}} = 220$ fs are the pulse durations, $\omega_{\text{pump}}/2\pi = \omega_{\text{probe}}/2\pi = 230$ THz are the carrier frequencies, $\tau$ is the time delay between the arrival of pump and probe pulses, and $E_{\text{probe},0}$ 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 [9, 32, 33]. 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 non-uniformity of $T_e$ and $T_{ph}$ in the transverse direction can be neglected. The initial temperature of electrons and phonons are set to be room temperature $T_0 = 300$ K. We also neglect the energy transfer from phonons to the surrounding due to the short time scale of interest.

We show below results for pump peak intensities of $I_0 = 5$ GW/cm$^2$ to 75 GW/cm$^2$. 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 photoluminescence, photoemission, and even for higher-order non-parabolicity of the band structure, multi-photon 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 non-thermal electron distribution on the ITO nonlinearity.

A. Temperature dynamics

For the purpose of characterizing the solution of the ANTH$^8$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 [34]. 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. Fig. 2(a) shows the dynamics of a rapid (pulse duration-limited) increase followed by electron cooling familiar from noble metals [19]. However, Fig. 2(a) also shows that unlike noble metals, the electron subsystem reaches extremely high electron temperatures (in particular, 8000 K), i.e., a considerable fraction of the Fermi temperature ($\sim 0.8 T_F$). Notably, since the pump pulse duration is much longer than the $e-e$ relaxation time $\tau_{e-e}$ (a few tens of fs, see Fig. 8(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 $C_e$ by $dT_e/dt = - G_{e-ph}(T_e)(T_e - T_{ph})/C_e(T_e)$, the electron-phonon energy coupling coefficient $G_{e-ph}$ and to the temperature difference between electrons and phonons [35]. Figs. 2(a)-(b) show 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) and the slope of $T_e(t)$ is more negative for higher $I_0$ (higher $T_e$). 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 [12]).

The drastic increase of $T_e$ for $I_0 \geq 50$ GW/cm$^2$ reveals a rather surprising result - the effective chemical potential $\mu$ can become negative. Specifically, Fig. 2(c) plots $\mu$ as a function of time for the case of $I_0 = 50$ GW/cm$^2$, showing that the chemical potential goes below zero after the pulse peaks at zero fs, and again recovers to positive values as $T_e$ decays. Consequently, the electron distribution becomes more flattened, as seen in Fig. 2(d), eventually becoming a distribution that is quite similar to the electron distribution in a semiconductor [25].

The cooling of the electrons results in an increase of $T_{ph}$, see Fig. 2(c). Fig. 2(c) also shows excellent agreement between the $T_{ph}$ dynamics in the ANTH@M and the TTM (Eq. (B12)). Moreover, the phonon temperature dynamics shows a decreasing growing rate. This, again, can be understood by the equation of energy balance Eq. (B12b), where the increase rate of $T_{ph}$ is given by $dT_{ph}/dt = G_{e-ph} \cdot (T_e - T_{ph})/C_{ph}$. As $t$ increases, the temperature difference between electrons and phonons becomes smaller, leading to a smaller $T_{ph}$ increase rate. With the same reasoning, the increase rate of $T_{ph}$ in ITO is much faster than that in noble metals because of the much higher $T_e$ reached [12].

B. Permittivity dynamics (nonlinear optical response)

In Fig. 3 we plot the dynamics of the ITO permittivity based on the ANTH@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 Fig. 2(a) and (b). This excellent match is remarkable in comparison to the phenomenological corrections introduced in a recent work [10]. In the latter, the permittivity calculations matched the experimental reflectance/transmittance [4] only if the damping coefficient $\eta$ was assumed to increase linearly with $T_e$, the $e-ph$ coupling coefficient was reduced a factor of 8, and the $e-ph$ energy transfer from the non-thermal part of the electron distribution is assumed to
FIG. 2. (Color online) (a) Electron temperature 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 [12]) and the effective temperatures obtained from the TTM (see Appendix B), respectively. (b) The decay rate of $T_e$ as a function of the electron temperature (the phonon temperature is set to be 300 K). (c) The (instantaneous) chemical potential (see details in [12]) as a function of time for pulse intensity $I_0 = 50$ GW/cm$^2$. (d) 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 (c). 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. (e) The same as Fig. 2(a) but for the phonon temperature $T_{ph}$.

be negligible. The self-consistent treatment presented here not only naturally connects the damping $\eta$ 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 $\varepsilon'$
12-fold increase for \( I_0 = 50 \text{ GW/cm}^2 \), which is a result of the initial proximity to the ENZ point [4 8 9 36]. One can see that the dynamics of \( \varepsilon' \) shares similar features with the time evolution of \( T_e \) (compare Figs. 2(a) and 3(a)). This is caused by the (nearly) linear decrease of the effective plasma frequency with \( T_e \) due to the non-parabolicity, which has been explained in many previous studies [3 5 8 37]. Therefore, by Fig. 2(b), the decay of the ITO permittivity is faster for higher \( T_e \). This indicates that the permittivity dynamics is not sufficient to explain the opposite behaviour of the experimentally-observed slowing of the reflection decay at high intensities in [9]. We return to this issue in Section III D.

In that regard, although ITO is a Drude material, the significant 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 [22 38]. Moreover, at the ENZ point, not only the field enhancement and absorption are strong, but also the relative change \( \varepsilon' (= \Delta \varepsilon'/\varepsilon') \) and thus the nonlinear optical response are maximized.

FIG. 3. (Color online) (a) The real part and (b) the imaginary part of the permittivity \([9]\) as a function of the time delay for the pulse intensity \( I_0 = 5 \text{ GW/cm}^2, 22 \text{ GW/cm}^2, 50 \text{ GW/cm}^2 \) and \( 75 \text{ GW/cm}^2 \) (the color used is the same as Fig. 2) at the carrier frequency. (c) The separation of the contributions to \( \varepsilon'' \) due to the change of \( \eta_{e-ph} \) (opened dots); and \( T_e \) (filled dots) for \( I_0 = 50 \text{ GW/cm}^2 \). The solid line is the same as in (b). (d) The separation of the effects of \( T_e \) on \( \varepsilon'' \) via the change of \( \eta_{e-e} \) (up-pointing triangles) and \( \omega_p^2 \) (down-pointing triangles). The orange solid line shows the change of \( \omega_p^2 \) (normalized to its value at 300 K) as a function of the time delay.

By contrast, \( \varepsilon'' \) shows a more complicated dynamics because it is proportional to the product of \( \omega_p^2 \) (Eq. (10)) and the damping term \( \eta \), and thus depends on both \( T_e \) and \( T_{ph} \). to see this, in Fig. 3(c), we separate the contribution due to the change of \( T_e \) (by setting \( T_{ph} = 300 \text{ K} \)) and of \( T_{ph} \) (setting \( T_e = 300 \text{ K} \)). Since \( T_{ph} \) affects \( \tau^{-1}_{e-ph} \) linearly [12] but does
not affect $\tau_{e-e}^{-1}$ nor $\omega_p$ (see Eq. (10)), the contribution due to $T_{ph}$ results in a slow increase of $\varepsilon''$ (see open dots in Fig. 3(c)). Fig. 3(c) also shows that the complicated dynamics of $\varepsilon''$ originates from the dependence on $T_e$ (see solid dots in Fig. 3(c)). In particular, when $T_e$ increases, $\tau_{e-e}^{-1}$ increases but $\omega_p$ decreases, see their opposite effects on the $\varepsilon''$ dynamics separated in Fig. 3(d). Thus, the $T_e$-dependence of $\varepsilon''$ can be roughly expressed as $\varepsilon'' \sim T_e^2(1 - \alpha T_e)$, indicating that $\varepsilon''$ increases for modest electron temperatures but decreases at higher ones. Therefore, when the pump peak intensity is relatively high (> 50 GW/cm$^2$ in Fig. 3(b) and (c)), the standard $\varepsilon''$ 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. 3(b) and (c)). That drop in $\varepsilon''$ is more pronounced as the illumination intensity is increased, compare the purple and green lines in Fig. 3(b). Importantly, the increase of the imaginary part of the ITO permittivity with time (and intensity) further indicates that the physics underlying the opto-thermal response of LEDD materials is similar to that of Drude metals (see [20, 21], 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 $\varepsilon'$ 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 Fig. 4(a)-(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 ($U_{abs}(I_0) = \int P_{abs}(t, I_0)dt$) from the self-consistent simulations via $T_{ph,max}(I_0) = T_0 + U_{abs}(I_0)/C_{ph}$. This is based on the fact that almost all the absorbed energy transfers to the phonon subsystem before leaking to out to the surrounding [39]. 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 $U_{pulse}$)) as shown in Fig. 4(c). It shows that at low illumination intensities, for which the permittivity is close to the ENZ point (see Fig. 3), most energy is absorbed (ENZ resonance). However, as the illumination intensity grows and the real part of the ITO permittivity drastically increases (see Fig. 3(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 only captured by a self-consistent simulation. This confirms that the decrease of the absorptivity does not originate from saturable absorption.

![Graphs](image)

FIG. 4. (Color online) (a) The maximum of the electron temperature as a function of the pulse peak intensity (extracted from Fig. 2(a)). (b) The same as (a) but for the phonon temperature $T_{ph}$. 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) is the same as (b) but is plotted in semilogx scale and is extended to 750 GW/cm$^2$. The open dots represent the estimated upper bound of the maximum $T_{ph}$ for $I_0$ up to 750 GW/cm$^2$ assuming that the absorptivity is constant for $I_0 > 75$ GW/cm$^2$. The dashed black line represents the damage threshold temperature (1900 K) of ITO reported in [9]. (c) The total absorbed pump pulse energy (normalized to the pump pulse energy) as a function of the pulse peak intensity (bottom x axis)/pump pulse energy density (top x axis).

In Fig. 4(b), we further extrapolate the data obtained from the simulation up to 750 GW/cm$^2$ (the opened dots) by assuming that the absorptivity remains the same for $I_0 > 75$ GW/cm$^2$, namely, $T_{ph,\text{max}}(I_0 > 75 \text{ GW/cm}^2) = T_0 + \frac{U_{\text{abs}}(I_0 = 75 \text{ GW/cm}^2)}{C_{ph}} \cdot \frac{I_0}{75 \text{ GW/cm}^2}$. The opened dots in Fig. 4(b) indicates that $T_{ph}$ reaches the melting point of ITO ($\sim 1900$ K) for $I_0 > 500$ GW/cm$^2$. This explains the high damage threshold of ITO observed experimentally in [9].

Finally, we note that the sublinearity of maximal $T_e$ vs. $I_0$ 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 ANTHεM to the experimental data [9], we 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 Section II B), the Ampère’s law and the Maxwell–Faraday equation for the probe pulse become

\[
\begin{aligned}
\nabla \times \mathbf{H}_{\text{probe}}(\mathbf{r}, t; \tau) &= -i \omega_{\text{probe}} \varepsilon_0 \varepsilon(t; \omega_{\text{pump}}, \omega_{\text{probe}}) \mathbf{E}_{\text{probe}}(\mathbf{r}, t; \tau), \\
\nabla \times \mathbf{E}_{\text{probe}}(\mathbf{r}, t; \tau) &= -i \omega_{\text{probe}} \mu_0 \mathbf{H}_{\text{probe}}(\mathbf{r}, t; \tau)
\end{aligned}
\]

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

\[
\lim_{q \to 0} \frac{2e^2}{\varepsilon_0 q^2} \int \frac{d^3k}{(2\pi)^3} \frac{f_{k+q}(t; \omega_{\text{pump}}) - f_k(t; \omega_{\text{pump}})}{\varepsilon_{k+q} - \varepsilon_k - \hbar \omega_{\text{probe}} - i\hbar(\eta_{k+q}(t; \omega_{\text{pump}}) + \eta_k(t; \omega_{\text{pump}}))}.
\]

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 Eq. (11) and (12) is correlated with the pump pulse envelope but not with any photon frequency (see Section II B). The temporal profile of the reflected probe pulse envelope is thus equal to the product of the Fresnel reflection coefficient and the incident probe pulse envelope, i.e., $E_{\text{probe, ref}}(t; \tau) = r(t; \omega_{\text{probe}})(E_{\text{probe, 0}}/2)e^{-2\ln(2(t-\tau)/\tau_{\text{probe}})^2}e^{-i\omega_{\text{probe}}(t-\tau)} + c.c.$. Notice that the Fresnel reflection coefficient is independent of the time-delay parameter since it originates from the (pump-induced) electron dynamics. Fig. 5 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 and that the frequency shift of the ENZ resonance 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, $R(\tau) = \frac{\int |E_{\text{probe, ref}}(t; \tau)|^2 dt}{\int |E_{\text{probe, inc}}(t; \tau)|^2 dt}$; it is, therefore, a function of the time delay. Here, $E_{\text{probe, ref}}(t; \tau) = r(t; \omega_{\text{probe}})(E_{\text{probe, 0}}/2)e^{-2\ln(2(t-\tau)/\tau_{\text{probe}})^2}e^{-i\omega_{\text{probe}}(t-\tau)}$ is the envelope of the reflected probe pulse. Fig. 6a shows that the reflection of the probe

16
pulse gradually increases with the pump peak intensity, even though at a decreasing rate. In addition, the time period for the change of the reflection reduces from its maximum to half this maximum is longer for higher pump peak intensities, see Fig. 6(b). This finding matches the observations in [9].

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. 5(c) and (d)) so that the dependence of the reflection change on the pump peak intensity becomes sublinear.

E. The response to shorter pulses

Since the duration of the pump pulses considered in previous subsections and in the experiments [2–6, 8–10] are much longer than the $\text{e} - \text{e}$ relaxation time, the dynamics of the temperatures and of the permittivity can be well approximated by the TTM, see also
FIG. 6. (Color online) (a) The probe reflection as a function of the pump-probe delay for different pump pulse peak intensities (the same colors are used as in Fig. 3). 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.

the interpretation of the experimental results [4] in [10]. 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 Section 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$\xi$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 non-thermal electrons on the nonlinear response of ITO. We also discuss the validity of using the third-order nonlinear optical susceptibility to describe the ultrafast optical nonlinearity. Fig. 7 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 3, but differ moderately 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 within 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 Fig. 7(a), (b), (d) and (e). In addition, comparing
FIG. 7. (Color online) The time dependence of (a) the electron distribution at $E = E_F + \hbar \omega_{\text{pump}}/2$, (b) the electron temperature, (c) the phonon temperature, (d) the real part and (e) the imaginary part of the ITO permittivity. The solid and dashed lines in (b) and (c) are obtained from the ANTHEM and from the TTM, respectively. In (d) and (e), the solid and dashed lines are calculated from Eq. (5) using the electron distribution obtained from the ANTHEM and the thermal distribution with the effective $T_e$, respectively. The purple, brown/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 $E = E_F + \hbar \omega_{\text{pump}}/2$. The inset in (b) is the zoom-in at $t = 50 - 250$ fs, showing the difference in $T_e$ between the ANTHEM and TTM ($\sim 200$ K) more clearly.

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 electron temperature (Fig. 7(b)), and thus, in a larger change of $\varepsilon'$ (Fig. 7(d)), and a more pronounced drop in the dynamics of $\varepsilon''$ (Fig. 7(e)).

Moreover, Fig. 7(b) and (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 ANTHEM. This occurs because the LTTM causes the change of $\varepsilon'$ to be slightly larger (see Figs. 7(d)),
resulting in 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 non-thermal electron distribution on the nonlinear optical response of ITO, we compare the ANTH\$M with the eTTM (see details in Appendix C). In contrast to the TTM, the eTTM accounts for the finite $e-e$ relaxation time thus allows the electron subsystem to be non-thermal. In this model, the electron subsystem is described by an instantaneous temperature that represents only the thermal part of the electron distribution. Fig. 9 shows that the results of the eTTM are in remarkable agreement with the ANTH\$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 $\varepsilon'$ is controlled by the pulse duration due to the account for the non-thermal distribution in the permittivity calculation. Moreover, if the non-thermal contribution is neglected when calculating the permittivity, the response time of $\varepsilon'$ will be controlled by the $e-e$ relaxation time, see Fig. 9(b). As a result, the absorption (Fig. 9(c)) and thus the maximum value of the instantaneous $T_e$ (Fig. 9(a)) are much higher compared with the results accounting for the non-thermal distribution in the permittivity calculation. This implies that the instantaneous rise time of $\varepsilon'$ is due to the change of the population, as claimed in [40]. Therefore, the account for the non-thermal 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 TTM, although the eTTM is more accurate, the result of the eTTM is only slightly better. 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 [10].

IV. CONCLUSIONS/DISCUSSION

We have presented a fully non-thermal model for the electron permittivity of LEDD materials, which is the first non-phenomenological model of such systems. Our work resolved several arguments and explained several experimental observations reported in LEDD material literature which were so far not understood, namely, the importance of momentum conservation in $e-ph$ interactions, the scaling of the collision rate with $T_e$, and the negligi-
ble energy transfer from non-thermal electrons to phonons. Our model also established the instantaneous nature of the pump-induced permittivity changes, and revealed the possibility of the chemical potential to transiently decrease dramatically, even to negative values.

The non-thermal (and the simpler thermal) permittivity model described above is 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., [4, 41], or later in [8, 10, 14]. However, it is known that for absorptive materials such as metals (and hence, LEDDs), the response is strongly non-local in time, such that the fitted values are strongly dependent on the pulse duration, see e.g., [42, 43]). As a particular example, Fig. 7 also shows that if the pump peak intensity is reduced to 22 GW/cm$^2$ (by a factor $\sim 8$) but the pulse duration remains 30 fs, the changes of the extracted $T_e$ and $\varepsilon'$ are only reduced by 50%. 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 Fig. 7(d)-(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 [4, 8, 44].

Thus, 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 non-equilibrium response that can be approximated by thermal models only for sufficiently long pulses.

This work is a starting point for the study of the nonlinear optical response of ITO at shorter wavelengths (e.g., [3]), nonlinear effects requiring very high nonlinearity such as bistability, and to study the response at higher intensities and shorter pulses. Our work may also be the starting point of the study of the nonlinear optical response of other TCOs such as TiN, ZrN [45] etc..

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
The electric displacement \( \mathbf{D}(t) \) is then related to the local electric field \( \mathbf{E}(t) \) by \( \varepsilon_0 \varepsilon_\infty \mathbf{E}(t) + \int_{-\infty}^{t} R(t,t') \mathbf{E}(t') dt' \),

\[
\mathbf{D}(t) = \varepsilon_0 \varepsilon_\infty \mathbf{E}(t) + \int_{-\infty}^{t} R(t,t') \mathbf{E}(t') dt',
\]

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

\[
R(t,t') = \lim_{q \to 0} \frac{2e^2}{i\hbar q^2} \int \frac{d^3k}{(2\pi)^3} e^{-[(\eta_k+q)(f(t'))+\eta_k(f(t')))/2+i(E_k+q-E_k)(t-t')/\hbar} (f_{k+q}(t') - f_k(t')).
\]

If the damping rate \( \eta_k \) is much faster than the change rate of the electron distribution, only the electron dynamics at \( t' \) nearby \( t \) contributes to the integral in Eq. \( \text{(A1)} \). 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 \( |\mathbf{E}(t)|^2 \)), the condition above is satisfied when \( |(df/dt)_{exc}| < \eta \). This requires the local field to be smaller than \( 2.7 \times 10^9 \text{ V/m} \). In this case, one can replace \( f_k(t') \) by \( f_k(t) \) and \( \eta_k(f(t')) \) by \( \eta_k(f(t)) \), 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 (\( \tau_{\text{pump}} \gg 2\pi/\omega_{\text{pump}} \)) and is also much longer than the damping time (\( \tau_{\text{pump}} \gg 1/\eta \)). The condition \( \tau_{\text{pump}} \gg 2\pi/\omega_{\text{pump}} \) allows us to write the electric field (and the electric displacement) as a product of slowly varying envelope \( (\mathbf{E}(t') \text{ or } \mathbf{D}(t')) \) and a rapidly varying phase factor with the carrier frequency, i.e., \( \mathbf{E}(t') = \mathbf{E}(t'e^{-i\omega_{\text{pump}}t'} + c.c. \text{ (} \mathbf{D}(t') = \mathbf{D}(t'e^{-i\omega_{\text{pump}}t'} + c.c.) \text{. After substitute } \mathbf{E}(t') \text{ back to Eq. } \text{(A1)}, the condition } \tau_{\text{pump}} \gg 1/\eta \text{ allows us to replace } \mathbf{E}(t') \text{ by } \mathbf{E}(t) \text{ so that the envelope function } \mathbf{E}(t) \text{ can be factored out of the integral. Then, Eq. } \text{(A1)} \text{ becomes}

\[
\mathbf{D}(t) = \varepsilon_0 \varepsilon_\infty \mathbf{E}(t) + (f_{k+q}(t) - f_k(t)) \mathbf{E}(t) - \lim_{q \to 0} \frac{2e^2}{i\hbar q^2} \int_{-\infty}^{t} dt' \int \frac{d^3k}{(2\pi)^3} e^{-[(\eta_k+q)(f(t'))+\eta_k(f(t')))/2+i(E_k+q-E_k)(t-t')/\hbar} e^{i\omega_{\text{pump}}(t-t')} \mathbf{E}(t'e^{-i\omega_{\text{pump}}t'} + c.c.).
\]

Eq. \( \text{(A3)} \) can be reorganized into the form

\[
\mathbf{D}(t) = \varepsilon_0 \varepsilon_\infty \mathbf{E}(t) + \varepsilon_0 \varepsilon_\infty \mathbf{E}(t) - \varepsilon_0 \varepsilon_\infty \mathbf{E}(t) e^{-i\omega_{\text{pump}}t} + \varepsilon_0 \varepsilon_\infty \mathbf{E}(t)e^{-i\omega_{\text{pump}}t} + c.c.
\]

When substituting the electric displacement field into the Ampèere’s law, the time derivative of electric displacement field becomes

\[
\frac{\partial \mathbf{D}(t)}{\partial t} = \varepsilon_0 \frac{\partial \mathbf{E}(t)}{\partial t} e^{-i\omega_{\text{pump}}t} - i\omega_{\text{pump}} \varepsilon_0 \mathbf{E}(t) e^{-i\omega_{\text{pump}}t} + c.c.
\]

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

Appendix B: Extended two-temperature model (eTTM) and two-temperature model (TTM)

Following the approach in [15, 24], the dynamics of the electron and phonon subsystems can be macroscopically described by deriving the extended two temperature model (eTTM) 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 non-thermal part, $f(\mathcal{E}) = f^T(\mathcal{E}, \mu(T_e), T_e) + f^{NT}(\mathcal{E})$, 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 non-thermal electron distribution $f^{NT}$ and its integral version (i.e., an equation for the non-thermal electron energy $U^{NT} = \int \mathcal{E} \rho_e(\mathcal{E}) f^{NT} d\mathcal{E}$),

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

(B1a)

$$\frac{dU^{NT}}{dt} = P_{\text{abs}}(t) + \left( \frac{dU^{NT}}{dt} \right)_{e-e} + \left( \frac{dU^{NT}}{dt} \right)_{e-ph},$$

(B1b)

as well as equations for the thermal electron energy,

$$C_e(T_e) \frac{dT_e}{dt} = \left( \frac{dU^T}{dt} \right)_{e-ph} - \left( \frac{dU^{NT}}{dt} \right)_{e-e},$$

(B2)

and phonon energy

$$C_{ph} \frac{dT_{ph}}{dt} = - \left( \frac{dU^T}{dt} \right)_{e-ph} - \left( \frac{dU^{NT}}{dt} \right)_{e-ph}.$$

(B3)
Here, $T_e$ is referred to as the instantaneous instantaneous electron temperature. The instantaneous is different from the extracted $T_e$ because the instantaneous $T_e$ has a longer rise time than the extracted $T_e$, as we shall see in Appendix C. $(\partial f/\partial t)_{e-e}$ is the electron-photon excitation term given by Eq. (2), while $(\partial f^{NT}/\partial t)_{e-e}$ and $(\partial f^{NT}/\partial t)_{e-ph}$ are, respectively, the rates of non-thermal electron relaxation via the $e-e$ and $e-ph$ interactions. $P_{\text{abs}}(t)$ is the absorbed power density given by Eq. (6), $(dU^{NT}/dt)_{e-e}$ is the non-thermal energy relaxation due to the $e-e$ interaction; $(dU^{NT}/dt)_{e-ph}$ and $(dU^{T}/dt)_{e-ph}$ are, respectively, the energy transfer from the non-thermal 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 non-thermal part of the electron distribution explicitly since the non-thermal 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,

$$
\left(\frac{\partial f^{NT}}{\partial t}\right)_{e-e} = \frac{f^{NT}}{\tau_{e-e}}, \tag{B4}
$$

where $\tau_{e-e}$ is the $e-e$ relaxation time. In many previous studies [4, 5, 10, 11, 49], $\tau_{e-e}$ was estimated using the Gurzhi formula [50–52],

$$
\frac{1}{\tau_{e-e}(T_e)} = \frac{\omega_{\text{pump}}^2}{4\pi^2\omega_p(T_e)} \left[ 1 + \left( \frac{2\pi k_B T_e}{\hbar \omega_{\text{pump}}} \right)^2 \right], \tag{B5}
$$

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 (B5) suggests that $\tau_{e-e}^{-1}$ exhibits a super-quadratic growth with $T_e$ (since the plasma frequency Eq. (10) decreases with $T_e$ [3, 5, 10, 11, 30]). In contrast, in the ANTH6M, we calculate the $e-e$ relaxation time from the functional derivative of the $e-e$ collision term. In particular, $\tau_{e-e}$ is shown [12] to be well described by the Fermi-liquid theory with a correction factor, i.e., $\tau_{e-e,k}^{-1} \sim (1 + 2C\mu(T_e)) (e_k^2 + (\mathcal{E}_k - \mu(T_e))^2)$. Due to the decrease of the chemical potential with $T_e$ [12] (see Fig. 2(c)-(d)), 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 $\tau_{e-e}^{-1}$ in [10].
FIG. 8. (Color online) (a) The $T_e$-dependence of the $e-e$ relaxation rate $\tau_{e-e}^{-1}$. The orange solid line represents the result calculated using Eq. (17) in [12] at $E = \mu(T_e) + \hbar\omega_{pump}/2$. The orange dashed line represents the result calculated using the Gurzhi formula (B5). (b) The rate of total energy transfer from the electron to the phonon subsystem Eq. (B7) (blue solid line). The green dashed line is the contribution from the thermal part of electron distribution (B8). The results correspond to case of $I_0 = 161$ GW/cm$^2$ and $\tau_{pump} = 30$ fs. (c) The electron energy dependence of $\Lambda_{e-ph}/\tau_{e-ph}$ (see Eq. (B9)) (green dashed line) and $\Lambda_{e-e}/\tau_{e-e}$ (orange dashed line) for the same case as (b) at $t = 0$ fs. The blue solid line is the non-equilibrium electron distribution before relaxation.

$$(\partial f^{NT}/\partial t)_{e-ph}$$ and $$(dU^{NT}/dt)_{e-ph}$$ were also treated using the relaxation time approximation in previous studies [4 5 11], i.e., $$(\partial f^{NT}/\partial t)_{e-ph} = -f^{NT}/\tau_{e-ph}$$ and $$(dU^{NT}/dt)_{e-ph} = -U^{NT}/\tau_{e-ph}$$, where $\tau_{e-ph}$ is the $e-ph$ relaxation time and was estimated using (e.g., [5])

$$\frac{1}{\tau_{e-ph}(T_{ph})} = \Gamma_0 \left[ \frac{2}{5} + \frac{4T_{ph}^5}{T_D^5} \int_0^{T_D/T_{ph}} \frac{z^4}{e^z - 1} dz \right], \tag{B6}$$

where $\Gamma_0 = 0.5342$ fs$^{-1}$, and $T_D = 900$ K is the Debye temperature. In the ANTHEM, $\tau_{e-ph}$ 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 $\tau_{e-ph}$ evaluated using the explicit form given in [12] is similar to that obtained from Eq. (B6) and is shorter than 10 fs, i.e., much shorter than the $e-e$ relaxation time. In this case, most of the non-thermal energy directly relaxes to the phonon subsystem after the generation of the non-thermal 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., [10]). This disagreement can be reconciled by comparing the total energy transfer rate from electrons to phonons (denoted as $(dU/dt)_{e-ph}$) with its contribution from the thermal electrons (i.e., $(dU^T/dt)_{e-ph}$). 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,

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

and

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

$(dU^{NT}/dt)_{e-ph}$ is then the difference between $(dU/dt)_{e-ph}$ and $(dU^T/dt)_{e-ph}$, i.e.,

$$
\left( \frac{dU^{NT}}{dt} \right)_{e-ph} = \left( \frac{dU}{dt} \right)_{e-ph} - \left( \frac{dU^T}{dt} \right)_{e-ph}.
$$

Our model shows that the total $e - ph$ energy transfer rate is dominated by the thermal part of the electrons, i.e., $(dU/dt)_{e-ph} \approx (dU^T/dt)_{e-ph}$, see Fig. 8(b). This means that the contribution from the non-thermal part of the electrons $(dU^{NT}/dt)_{e-ph}$ is in fact negligible (see also [10, 15, 24]), confirming the ad hoc neglect of $(dU^{NT}/dt)_{e-ph}$ done in [10].

To have a deeper understanding of this, we cast the $e - ph$ collision term into a RTA form that enforces number conservation (by adding the aka Lorentz term, see [24]), namely,

$$
\left( \frac{\partial f}{\partial t} \right)_{e-ph} = -\tau_{e-ph}^{-1} \cdot f + \Lambda_{e-ph}.
$$

Thus, $\tau_{e-ph} \cdot \Lambda_{e-ph}$ can be interpreted as the target distribution which the non-equilibrium 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. Fig. 8(c) shows that the target distribution clearly differs from a thermal distribution and differs from the distribution before relaxation only slightly. 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. 8(c). Moreover, the $e - e$ Lorentz term is two 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 ($\left| \left( \frac{\partial f}{\partial t} \right)_{e-e} \right| \gg \left| \left( \frac{\partial f}{\partial t} \right)_{e-ph} \right|$, see Fig. 2 in [12]) although $\tau_{e-e}^{-1} < \tau_{e-ph}^{-1}$. The comparison above further indicates that the $e - ph$ interaction is mainly responsible for transferring
the thermal energy of electrons to phonon, instead of thermalizing the non-equilibrium electron distribution. Therefore, the $e-ph$ interaction violates the RTA assumption, thus it is improper to apply the relaxation time approximation for the $e-ph$ relaxation of the non-thermal electron energy. This conclusion is also valid for noble metals.

The energy transfer rate from the thermal electrons to the phonons $(dU_T/dt)_{e-ph}$ (Eq. (B8)) has been shown in [12] to be proportional to the temperature difference between electrons and phonons, namely, $(dU_T/dt)_{e-ph} = -G_{e-ph}(T_e - T_{ph})$, where $G_{e-ph}$ 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 [5, 10, 11]. However, we showed in [12] that this approach overestimates the correct value by 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 [10] to fit the experimental data.

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

$$C_e(T_e) = \int \rho_e(\mathcal{E}) \frac{\partial f^T(\mathcal{E}, \mu(T_e), T_e)}{\partial T_e} d\mathcal{E}, \quad (B10)$$

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

$$\int \rho_e(\mathcal{E}) f^T(\mathcal{E}, \mu(T_e), T_e) d\mathcal{E} = \int \rho_e(\mathcal{E}) f^T(\mathcal{E}, E_F, 0 K) d\mathcal{E}. \quad (B11)$$

The phonon heat capacity is taken to be $C_{ph} = 2.54 \times 10^6 \text{ J} \cdot \text{m}^{-3} \cdot \text{K}^{-1}$ [53].

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 (TTM) [54],

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

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

Here, $T_e$ is referred to as the effective electron temperature. In general, the effective $T_e$ is different from the extracted and instantaneous $T_e$ since thermalization is assumed to be instantaneous in Eq. (B12).

As shown in this work and others [10, 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 $I_0 = 161 \text{ GW/cm}^2$ studied in Section III E to demonstrate the importance of accounting for the non-thermal part of the electron distribution in the permittivity calculation when self-consistently solving the eTTM (Eqs. (B1)-(B3)) coupled with Maxwell’s equations (4). 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 $f(E) = f^T(E, \mu(T_e), T_e) + f^{NT}(E)$. In the second one, we neglect $f^{NT}(E)$ and evaluate the ITO permittivity using only the thermal distribution $f(E, \mu(T_e), T_e)$.

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. 9(a). In contrast, the rise time of $\varepsilon'$ (and thus the decrease of absorption) are controlled by the pulse duration, showing remarkable agreement with the results obtained from our ANTH\$M (see Fig. 9(b) and (c)).

However, if one neglects the non-thermal part of the electron distribution and simply evaluates the ITO permittivity using a thermal distribution with $T_e$, the rise time of the permittivity (and the decrease of absorption) will be the same as $T_e$ thus will be controlled
by the $e-e$ relaxation time instead of the pulse duration, see the dotted lines in Fig. 9. 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 non-thermal distribution in the permittivity calculation. This comparison clearly shows that the non-thermal electron distribution is essential to properly capture the nonlinear response of ITO.

The need to account for the non-thermal 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 non-thermal electron distribution was assumed to have the same profile as the photon excitation term, i.e., $f_{NT}(\mathcal{E}, t) = A(t)\left[f_T(\mathcal{E} - \hbar\omega_{\text{pump}}) \left(1 - f_T(\mathcal{E})\right) - f_T(\mathcal{E} + \hbar\omega_{\text{pump}}) \right]$, where $A(t)$ is determined from the non-thermal energy $U_{NT}(t)$ [38, 49, 55]. 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 non-thermal model (ANTHEM).

[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, Advanced Materials 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. Photonics 10, 267 (2016).
[4] M. Z. Alam, I. D. Leon, and R. W. Boyd, Large optical nonlinearity of indium tin oxide in its epsilon-near-zero region, Science 116, 795 (2016).
[5] 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. Photonics 12, 79 (2018).
[6] 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).
[7] 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).

[8] 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).

[9] 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. Applied (2022).

[10] J. Baxter, A. Pérez-Casanova, L. Cortes-Herrera, A. C. Lesina, I. D. Leon, and L. Ramunno, Understanding the nonlinear optical response of epsilon near zero materials in the time-domain, arXiv:2110.14806 (2021).

[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. (2022).

[12] S. Sarkar, I. W. Un, and Y. Sivan, The electronic and thermal response of low electron density drude materials to ultrafast optical illumination, arXiv:2207.14528v1 [physics.optics] (2022).

[13] In particular, the dependence of the $e^{-e}$ collision rate on the energy difference of the colliding electrons was fixed to the absorbed photon frequency, thus, overestimating the thermalization rate; the $e^{-e}$ collision rate is anyhow quite fast, this is likely to have a small effect on the results.

[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. Applied 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, 2nd ed. (Academic Press, San Diego, 2003).

[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] 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).

[20] 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).

21] 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).

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

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

24] Y. Dubi and Y. Sivan, “hot electrons” in metallic nanostructures - non-thermal carriers or heating?, Light: Sci. Appl. 8, 89 (2019).

25] 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).

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

27] N. W. Ashcroft and N. D. Mermin, Solid state physics (Brooks/Cole, 1976).

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

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

30] 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).

31] In this case, the factor $1/(\omega_{\text{pump}} + i\eta)^2$ can be factored out of the integral, so that one can write $\varepsilon = \varepsilon_{\infty} - \omega_{\text{p}}^2/(\omega + i\eta)^2$, where

$$\omega_{\text{p}}^2[f(\mathcal{E})] = \frac{e^2}{\varepsilon_0 m^*} \int p_\varepsilon(\mathcal{E}) f(\mathcal{E}) \left(\frac{1 + 8C\mathcal{E}(1 + C\mathcal{E})/3}{(1 + 4C\mathcal{E}(1 + C\mathcal{E}))^{3/2}}\right) d\mathcal{E}$$

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

reproducing the widely-used formula for the electron temperature dependent plasma frequency [3, 5, 10, 11, 30] for non-parabolic band dispersion.

32] T. G. Mackay and A. Lakhtakia, The Transfer-Matrix Method in Electromagnetics and Optics (2020).

33] M. Born, E. Wolf, A. B. Bhatia, P. C. Clemmow, D. Gabor, A. R. Stokes, A. M. Taylor, P. A.
The total energy of the electron system $U$ is the first moment of the electron distribution, i.e.,

$$U = \int E \rho_e(E)f(E)dE.$$  

The electron temperature associated with an electron distribution $f(E)$ is then determined by

$$\int E \rho_e(E)f^T(E,\mu(T_e),T_e)dE = \int E \rho_e(E)f(E)dE,$$

where $\mu(T_e)$ is the chemical potential given by Eq. (B11), see details in [12].

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 Fig. 2(b).

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 $\epsilon$-near-zero materials, Phys. Rev. Lett. 116, 233901 (2016).

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. Photonics 11, 390 (2017).

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 Photonics 2, 521 (2015).

This is because electrons stop transferring their energy once $T_e$ and $T_{ph}$ are the same. In 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 $C_e \ll C_{ph}$).

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, https://arxiv.org/ftp/arxiv/papers/2206/2206.04362.pdf.

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).
[42] 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).

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

[44] 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, Nature Communications 11, 2180 (2020).

[45] 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).

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

[47] L. D. Landau, J. Bell, M. Kearsley, L. Pitaevskii, E. Lifshitz, and J. Sykes, Electrodynamics of continuous media, Vol. 8 (elsevier, 2013).

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

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

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

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

[52] C. Voisin, N. Del Fatti, D. Christofilos, and F. Vallée, Ultrafast electron dynamics and optical nonlinearities in metal nanoparticles, The Journal of Physical Chemistry B 105, 2264 (2001).

[53] 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).

[54] 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).
[55] C. K. Sun, F. Vallée, L. H. Acioì, E. P. Ippen, and J. G. Fujimoto, Femtosecond tunable measurement of electron thermalization in gold, Phys. Rev. B 50, 15337 (1994).