nature nanotechnology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41565-019-0515-x

In the format provided by the authors and unedited.

Ultrafast coherent nonlinear nanooptics and nanoimaging of graphene

Tao Jiang[1,2], Vasily Kravtsov[1,2,3], Mikhail Tokman[4], Alexey Belyanin[5*] and Markus B. Raschke[1,2*]

[1]Department of Physics, Department of Chemistry and JILA, University of Colorado, Boulder, CO, USA. [2]Center for Experiments on Quantum Materials, University of Colorado, Boulder, CO, USA. [3]ITMO University, Saint Petersburg, Russia. [4]Institute of Applied Physics, Russian Academy of Sciences, Nizhny Novgorod, Russia. [5]Department of Physics and Astronomy, Texas A&M University, College Station, TX, USA. *e-mail: belyanin@physics.tamu.edu; markus.raschke@colorado.edu

Supporting Information for

Ultrafast coherent nonlinear nanooptics and nanoimaging of graphene

Tao Jiang, Vasily Kravtsov, Mikhail Tokman, Alexey Belyanin,[*] and Markus B. Raschke[†]

Supplementary Note 1: Near-field graphene FWM

The near-negligible FWM response of the tip under the tip-perpendicular excitation condition, the weak near-field enhancement between the tip and sample, and the generally inefficient tip scattering in the absence of the tip-parallel antenna effect, together, highlight the large efficiency of the nano-confined graphene FWM excited from an estimated near-field interaction area as small as ~(10 nm)$^2\pi$, which (as determined by the tip radius confined local field distribution), corresponds to as few as 10^4 atoms. Despite the nano-localized excitation, the Doppler broadening of the resulting FWM polarization distribution leads to delocalized radiative emission from an extended >100 nm diameter area.

Supplementary Note 2: Distinct enhancement at edges in far-field and near-field FWM imaging

Near-field graphene FWM imaging exhibits an enhanced FWM signal along edges (Fig. 1e and Fig. 2-3). However, this feature is absent in corresponding far-field control experiments. Figure S1a shows far-field FWM imaging of the same graphene sample presented in Fig. 2b. The edges show no distinct difference in FWM intensity compared to the internal sheet region. Figure. S1b-c show this contrasting behaviour on another sample in direct comparison, with FWM enhancement at the edge in near-field (Fig. S1b), but not in far-field imaging (Fig. S1c). Despite the ~400 nm spatial resolution limit of our far-field imaging, a 200-300 nm edge enhancement would still manifest itself in a spatially convoluted enhanced far-field edge signal if present. This indicates that the Doppler broadening associated with high momentum states of the interacting field is only present in near-field FWM.

In the vicinity of the graphene edge, the broken symmetry of the edge lifts the destructive interference of in-plane FWM polarization and thus gives rise to stronger FWM signals in near-field imaging. However, this effect does not exist in far-field FWM since the excitation field does not lead to an in-plane destructive FWM polarization density.

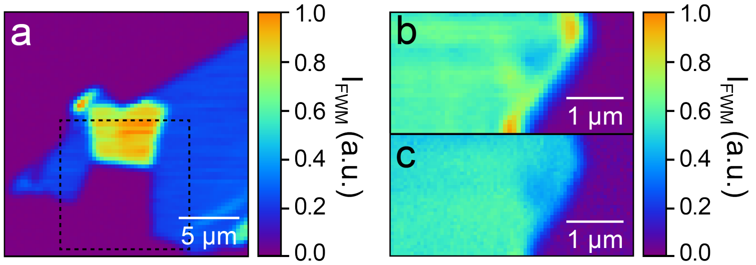

Figure S1. **Far-field (a,c) and near-field (b) graphene FWM imaging.** (a) Far-field FWM imaging of the same graphene sample area (dashed line box) as shown in Fig. 2b. (b-c) Near-field (b) and corresponding far-field (c) FWM imaging of another graphene sample. No FWM enhancement at graphene edges is observed in far-field imaging.

Supplementary Note 3: FWM heterogeneity and spatial resolution of FWM nanoimaging

Structural heterogeneities disrupt in-plane translational invariance and spatially localize the near-field interaction, which gives rise to increased localization of the FWM signal to the ultimate tip-radius near-field confinement limit. Figure S2 shows corresponding examples of increasing spatial localization with increasing perturbation of translational invariance in the graphene sheet, from large scale (~100's nm) standard heterogeneities (Fig. S2a), to finer scale wrinkles (~10's nm, Fig. S2c) in graphene. In contrast, simple roughness features in Au give rise to a tip apex size related highest spatial localization of the FWM near-field response (~15 nm, Fig. S2e).

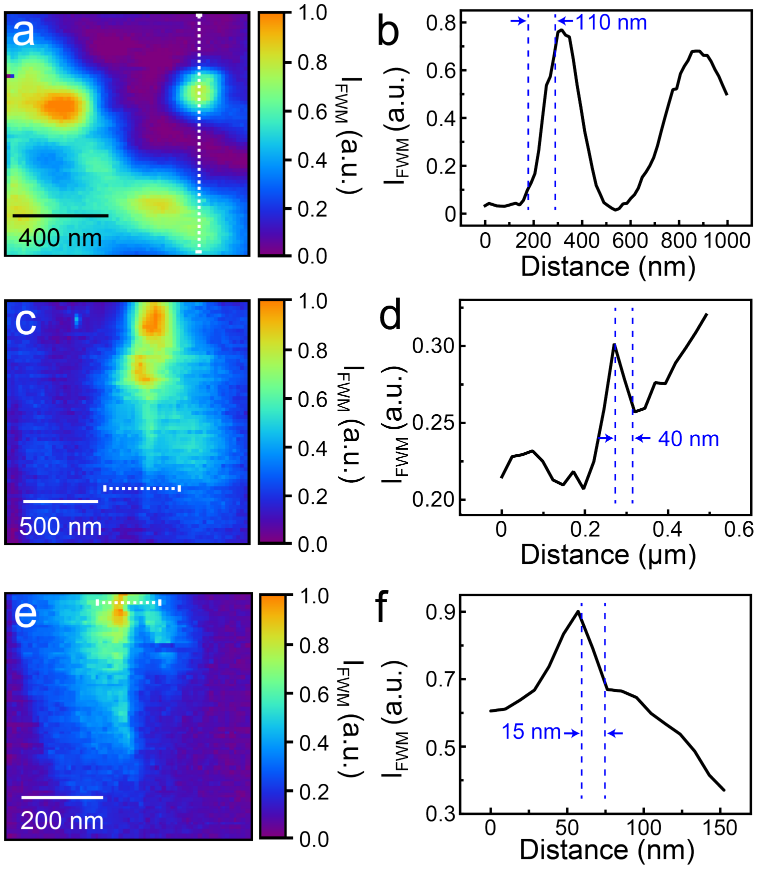

Figure S2. **FWM nanoimaging of structural heterogeneities.** FWM nanoimaging of graphene heterogeneities (a), wrinkled graphene with strains (c), and a rough Au surface (e), showing spatial confinement to ~110 nm (b), ~40 nm (d), and ~15 nm (f), respectively, extracted from the FWM signal variation along the white dashed lines, and defined by the signal rise from 10% to 90%.

Supplementary Note 4: Circumferential edge-parallel FWM

At internal and external boundaries, the spatial FWM source polarization is found to be oriented parallel with respect to the edges as seen in Fig. 3a-c. As a result of the far-field radiation from in-plane dipoles, we expect a strong FWM signal with horizontal polarization along the y-axis when detected under s polarization (Fig. 3b). For p polarization, due to the inefficient collection geometry for the in-plane electric dipole radiation, an only $\sim$7 times smaller intensity of the FWM signal polarized along the x-axis is observed (see normalized image in Fig. 3c). Figure 3b-c show the edge contrast reversal of the polarized FWM images. Such a circumferential edge-parallel FWM response mainly results from the radially asymmetric in-plane FWM current distribution at graphene edges, which further enhances the FWM signal due to the weaker cancellation effect of in-plane dipoles.

Supplementary Note 5: Resonant electronic transitions and large momenta compensation from tip.

According to the microscopic expressions for $\chi^{(3)}$ for graphene difference-frequency mixing [1, 2], there are five possible resonant transition pathways in total (Fig. S3a-e), yet only (c-e) dominate the FWM process and show a $\delta\omega = \omega_1 - \omega_2$ frequency detuning dependence. However, in-plane momentum is required to resonantly connect all photon interactions, with example shown in (f) for the pathway (e). Based on the linear energy dispersion $k = \omega/v_F$, for excitation condition $\hbar\delta\omega \sim 40$ meV, an in-plane momentum difference $k_{//} \sim 63\ \mu\text{m}^{-1}$ would be required for resonant interaction. In near-field induced FWM processes, the large plasmon wave vectors $q \sim \pi/R$ generated at the tip apex can reach beyond $\sim 300\ \mu\text{m}^{-1}$ for typical tip radii of $R \sim 10$ nm [3], which compensates the momentum difference of nonresonant transitions and makes graphene FWM even more efficient.

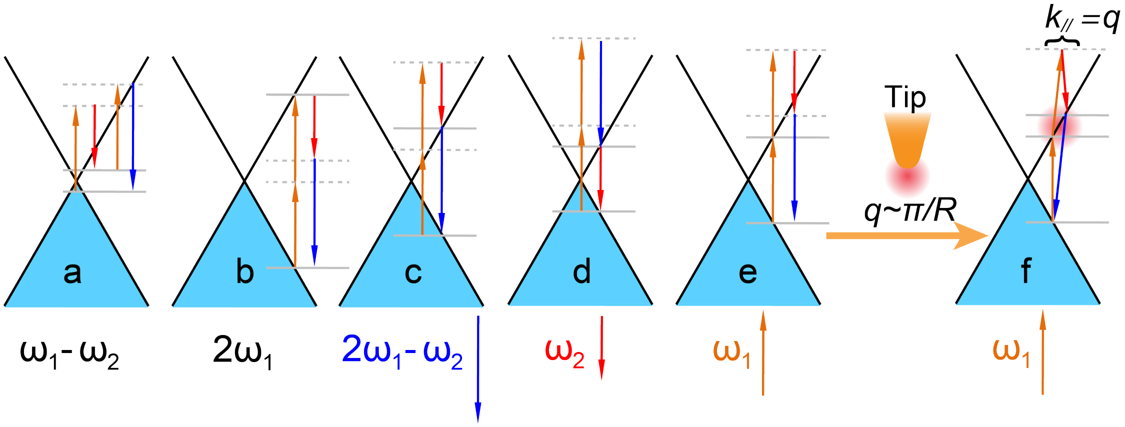

Figure S3. **Resonant transitions via near-field momenta at tip apex.** The left five panels (a-e) show the resonant pathways in graphene FWM processes with each corresponding photon resonant energy indicated below. The right panel (f) shows the tip near-field wave vector momentum matching effect for pathway (e).

We estimate the SPP adiabatic efficiency gain and associated near-field momenta using the adiabatic nanofocusing model. The result is calculated from analytical expressions for SPP propagation on a cone following established procedures [4–7], so neither mesh dimensions nor volume affect the result. The tip material is gold and no boundaries are assumed. Figure S4a-b show the spatial distribution of field enhancement for in-plane (tip-perpendicular, x) and out-of-plane (tip-parallel, z) electric field (E-field) components E_x and E_z, calculated for a gold tip with an apical angle of 6°. In the vicinity of a 10 nm tip apex, optical fields are enhanced 10-20 times, as seen from the corresponding E-field profiles in x and z directions (Fig. S4c-d). Due to the adiabatic compression of SPPs, with effective refractive index n_{eff} diverging with decreasing tip radius R as $\sim 1/R$, SPP wave vector q_z at $R = 10$ nm can be estimated as $q_z \sim 23\ \mu\text{m}^{-1}$, about 3 times higher than that in free space (see Fig. S4e).

We further estimate the distribution of k-vectors near tip apex via Fourier transform of the E-field spatial distribution. As shown in Fig. S4f, the distribution of q_z (black) is relatively broad due to strong near-field gradients at the tip apex, easily reaching the momentum of $\sim 63\ \mu\text{m}^{-1}$ (blue dashed line) required for the process in Fig. S3f. Distribution of in-plane momenta q_x is also broad (red), due to the field confinement in the transverse direction. We estimate the efficiency of the mechanism to generate in-plane momenta through the ratio of the two distributions at $q = 63\ \mu\text{m}^{-1}$ as $E(q_x)/E(q_z) \sim 0.25$. We note, however, that such efficiency can depend strongly on the nanoscale details and asymmetry of the tip apex.

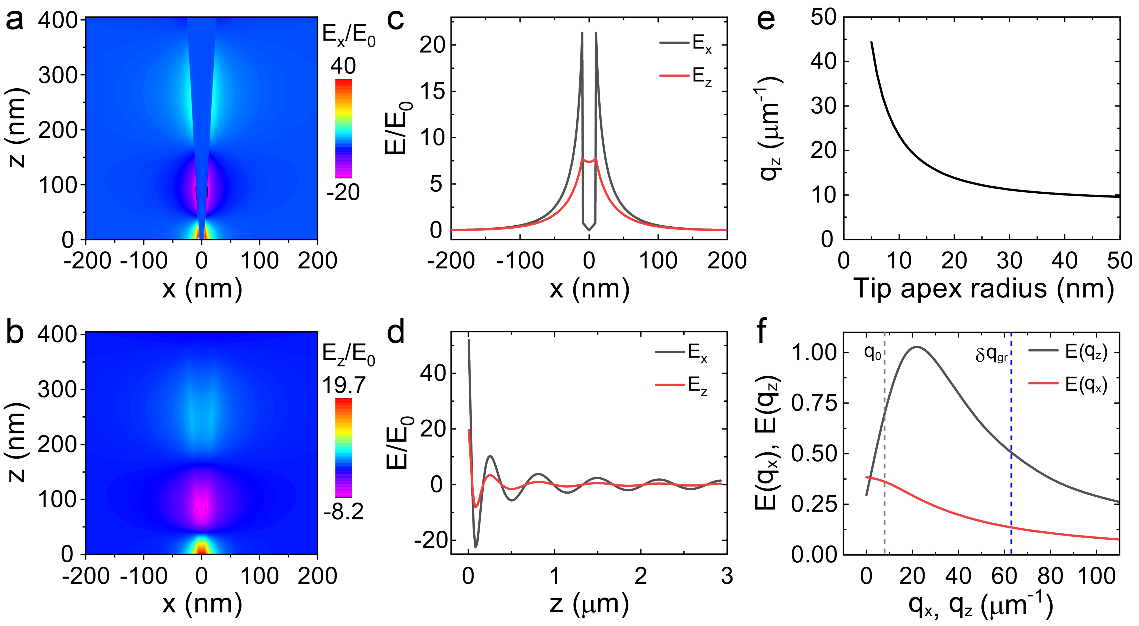

Figure S4. **Near-field distribution and associated enhanced wave vectors.** In-plane (a) transverse E-field distribution (E$_x$) and out-of-plane (b) longitudinal E-field distribution (E$_z$). (c) Profiles of E-field along x-coordinate (in-plane). (d) Profiles of E-field along the tip surface. (e) Out-of-plane component of SPP k-vector as a function of tip radius. (f) Fourier transforms of (c) and (d) yielding E-field distribution in k-space. Gray dashed line indicates the low free space photon momentum, blue dashed line indicates momentum required for the nonlinear process in Fig. S3f and enhanced in near-field.

Supplementary Note 6: FWM dynamics in graphene measured by a non-resonant tip

Figure S5a shows a non-resonant instantaneous FWM response of a tip with simulation with $T_2 = 0$ fs matching the FWM autocorrelation trace [6]. In addition to the FWM dynamics of the graphene edge (Fig. 5b), the FWM dynamics of the graphene sheet (left red square in Fig. 5a) is shown in Fig. S5b, which is described by a dephasing time of $T_2 = (6 \pm 1)$ fs, similar to the dynamics resolved at the edge. Note that the two delays $\tau = 5.6$ fs and 11.2 fs in Fig. 5 and Fig. S5b were set slightly offset from the exact peak positions at 5.5 fs and 11 fs, respectively, but the corresponding FWM data (blue circles) fall well onto the simulated FWM autocorrelation trace with $T_2 = (6 \pm 1)$ fs.

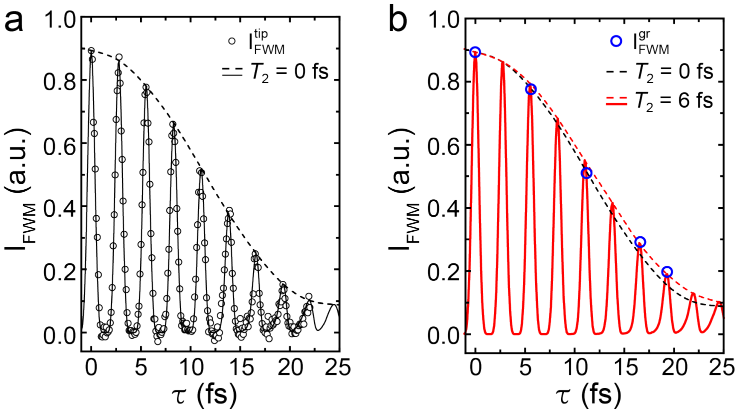

Figure S5. **FWM dynamics in tip and graphene.** (a) Spectrally integrated tip FWM autocorrelation trace (black circles) and simulated FWM response with $T_2 = 0$ fs (black line, together with black dashed envelope). (b) Extracted FWM dynamics (blue circles) of graphene sheet area indicated by left red square in Fig. 5a for the five delays $\tau = 0$ fs, 5.6 fs, 11.2 fs, 16.6 fs, and 19.3 fs, showing a finite decoherence time of $T_2 = (6 \pm 1)$ fs (red line, together with red dashed envelope) compared to a simulated instantaneous response (black trace envelope only).

Supplementary Note 7: Femtosecond spatio-temporal FWM imaging

In addition to the data shown in Fig. 5, we explore the FWM decoherent behaviour of another graphene sample with spatial FWM heterogeneities (Fig. S6). Visualizing the FWM decoherence behaviour across the whole graphene sample, there is no discernible spatial variation in T_2, even for the two extreme points of spatially homogeneous region (B) and large local strain or wrinkle associated localization with enhanced FWM (A) (Fig. S7).

Graphene FWM dynamics at points (A) and (B) also show a similar decay behaviour compared to the tip FWM. Different from the non-resonant tip used in Fig. 5 and Fig. S5, the tip used in this imaging experiment exhibits a resonant FWM response with $T_2 = (5 \pm 1)$ fs. The finite tip response does not allow to discernibly quantify T_2 of graphene in this

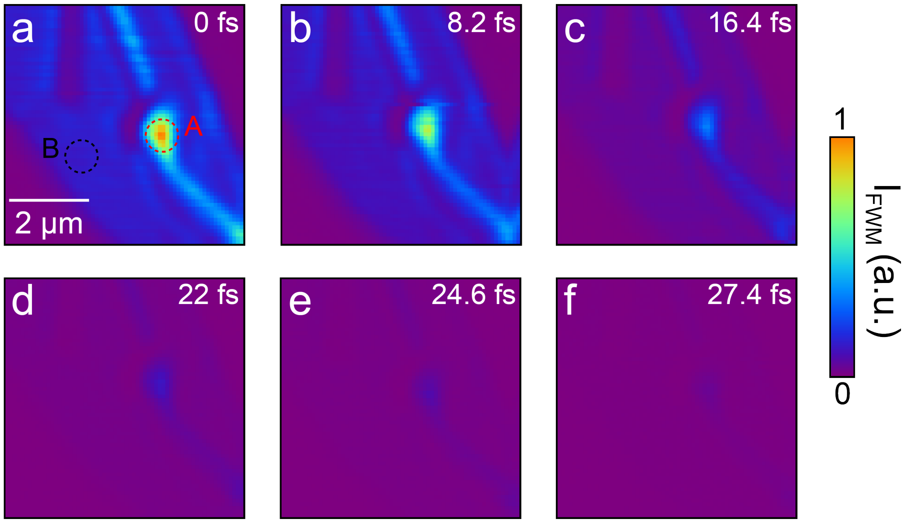

Figure S6. **Ultrafast FWM imaging.** Femtosecond FWM spatio-temporal imaging of the same graphene region with two-pulse excitation at different inter-pulse delays. The signal is strongest for zero delay decaying almost completely within ~ 20 fs. Animated GIF available online.

experiment. However, the observation of the extremely fast decay is still consistent with the ultrafast $T_2 = (6 \pm 1)$ fs extracted in Fig. 5.

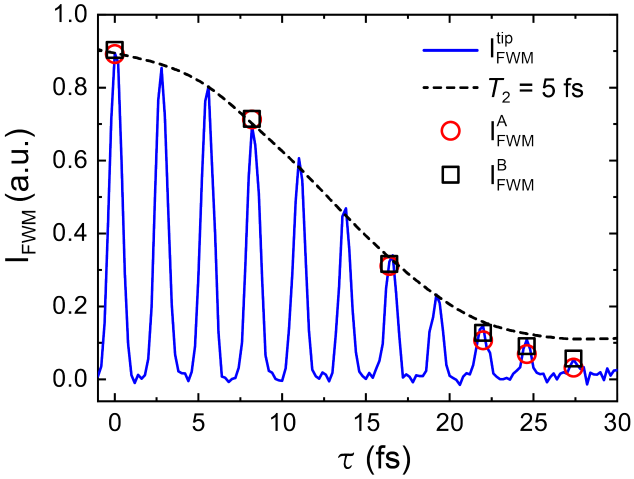

Figure S7. **Coherent FWM dynamics.** Extracted FWM intensity of points A (red) and B (black) for the six delay times, shows no discernible spatial variation in T_2, corresponding to an upper limit of $T_2 \leq 6$ fs.

Supplementary Note 8: Numerical aperture dependence of enhanced FWM edge width

We rule out a far-field artifact by varying the FWM detection numerical aperture, verifying the pure near-field signature of the emission (Fig. S8). This verifies the presence of a new effect giving rise to the spatial delocalization, which we attribute to a nonlocal contribution to $\chi^{(3)}$ in near-field FWM of graphene.

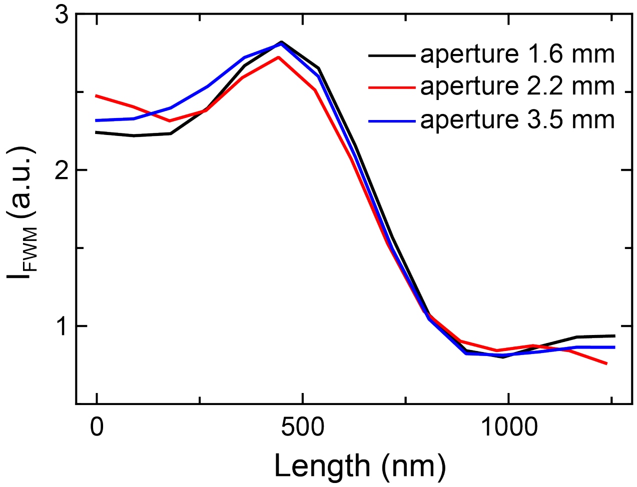

Figure S8. **Numerical aperture dependence of the width of the enhanced FWM signal at the graphene edge.** The enhanced FWM widths at a graphene edge for three different aperture sizes for the FWM signal beam. 95% of FWM signal pass the aperture with a size of 3.5 mm. The line width variation from 250 ± 3 nm to 280 ± 4 nm is much smaller than the spatial resolution of far-field emission response (> 400 nm) while constricting the numerical aperture, indicating that the underlying spatial delocalization effect (nonlocal FWM process) is intrinsic to the near-field FWM of graphene and not a far-field artifact.

Supplementary Note 9: AFM imaging of graphene.

The corresponding AFM images (Fig. S9) for the graphene samples in Fig. 1e and Fig. 2b verify the high quality of graphene edges, exhibiting a well defined outline of the graphene flakes with sharp edges and only minimal edge roughness on at most few nm length scale. With the AFM images, we can rule out the possible influences of defects, folds, and other contaminations near the edges responsible for the delocalized FWM response at the edges.

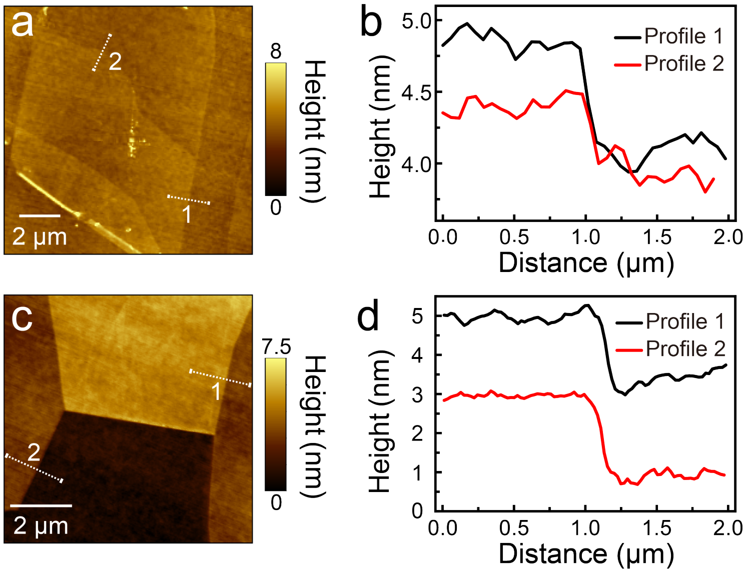

Figure S9. **High resolution AFM images of graphene samples.** AFM images (a) and (c) show the topography of the graphene samples studied in Fig. 1e and Fig. 2b, respectively. The line profiles (b) and (d) extracted from the height variation along the white dashed lines in the AFM images show that the edges are clean, flat and without folds.

Supplementary Note 10: Micro-Raman mapping of graphene.

Raman spectroscopy provides information on the physical, chemical, and electronic properties of graphene. Micro-Raman mapping (Fig. S10a-b) is performed to analyze the edge structure, defects and folds. A homogeneous G-mode Raman (I^G_{Raman}) and absence of D-mode Raman signal (I^D_{Raman}) verify the high quality of graphene, and reveals no correlation between lattice structural related heterogeneities and FWM broadening. Raman spectra (Fig. S10c-d) are used to determine the graphene layer thickness, and the assigned layer numbers are shown in Fig. 1e. We also verified that the unfolded area in Fig. 2b is trilayer.

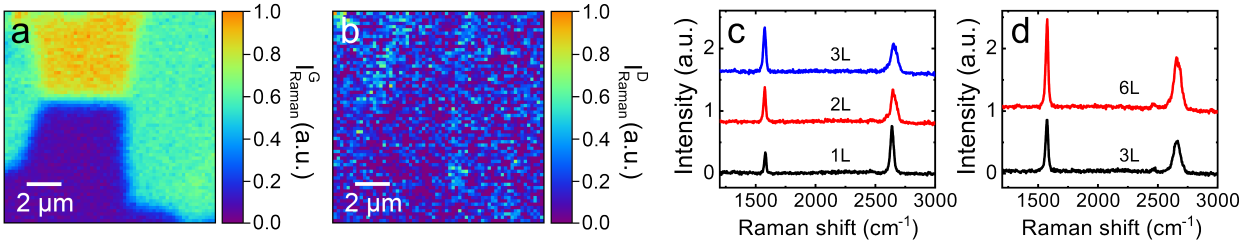

Figure S10. **Micro-Raman mapping of G-mode (a) and D-mode (b) of graphene in Fig. 2b, and the Raman spectra of graphene in Fig. 1e (c) and Fig. 2b (d).** G-mode Raman signal observed in graphene flake with negligible D-mode response.

Supplementary Note 11: FWM imaging of graphene on different substrates.

As a widely used substrate for graphene, SiO_2/Si has higher surface quality, lower roughness and facilitates the exfoliation of graphene with high quality. Thus we used SiO_2/Si as the substrate for the majority of the experiments. The graphene samples prepared on SiO_2/Si are further required to demonstrate the doping dependence. For comparison and as control experiments, we also performed near-field FWM imaging of graphene exfoliated on gold substrates, and ferroelectric nanorod arrays, of composition $PbZr_{0.52}Ti_{0.48}O_3$ (PZT) (Fig. S11). As can be seen, the FWM delocalization at the graphene edges is substrate independent, and shows that edge enhancement and spatial broadening are universal and robust features.

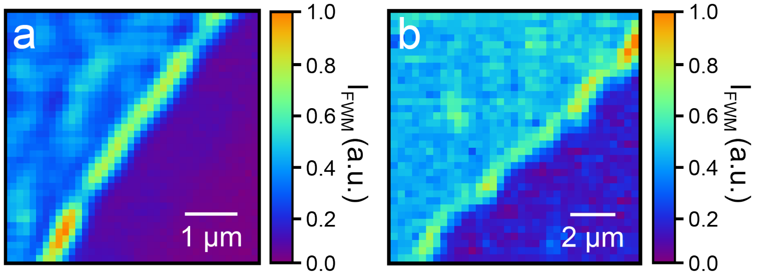

Figure S11. **FWM imaging of graphene on PZT (a) and gold substrates (b).** The near-field FWM images of graphene samples exfoliated on PZT and flat template-stripped gold substrates, show the universal FWM edge enhancement and spatial broadening delocalization.

Supplementary Note 12: Gate voltage dependence of FWM nano-imaging of graphene FET device.

The nonlinear wave mixing has been demonstrated to excite surface plasmons in graphene, as proposed in [8] and studied experimentally in [9]. The near-field FWM process may be mediated by surface plasmon generation, affecting the near-field FWM efficiency and edge enhancement. Since the Dirac plasmon in graphene is gate-tunable [10–13], a graphene field effect transistor (FET) provides the ability to study the effect of Dirac plasmons on the FWM process. On the other hand, the presence of oxygen defects in SiO_2 substrate results in hole-doping of graphene. It is also necessary to investigate the effect of defects on Fermi level and in turn on the FWM process. Thus we employed a graphene FET device to study the possible gate dependence of the near-field FWM efficiency (Fig. S12).

For graphene FET devices we use a SiO_2 layer as the gating material, with typical dielectric constant ~4, thickness ~90 nm, hence a gate capacitance ~4×10^{-2} μF/cm^2. We applied a back gate voltage V_g from -20 V to 20 V, studied the FWM integrated intensity across the edge, and FWM intensity variation at 3 locations: on graphene grain (black), on edge (red), and on substrate (blue). We did not observe a noticeable change of the FWM signal efficiency or edge enhancement as a function of Fermi level at any location. Taking into account the charge neutral point V_{CNP} = 18.5 V due to the intrinsic doping in this device, the highest Fermi level = 0.36 eV with hole doping was achieved with V_g - V_{CNP} = -38.5 V. In order to effectively tune the FWM response by blocking the resonant pathways [2] as illustrated in Fig. S3, a Fermi level larger than half of the excitation photon energies (~860 nm) is required which means V_g - V_{CNP} would have to reach -160 V, a value too large to be sustained by the thin SiO_2 gate layer. Alternative, gating methods such as ionic liquid gating to shift the Fermi level over a wide range is not suitable for near-field measurements. The near-field gate dependence of the FWM clearly indicates that the nonlocal effect is independent of gating for low doping. Albeit a negative result, with adiabatic nanofocused FWM nano-imaging being already a complex process, this example provides a perspective for possible future nano-FWM experiments on active devices.

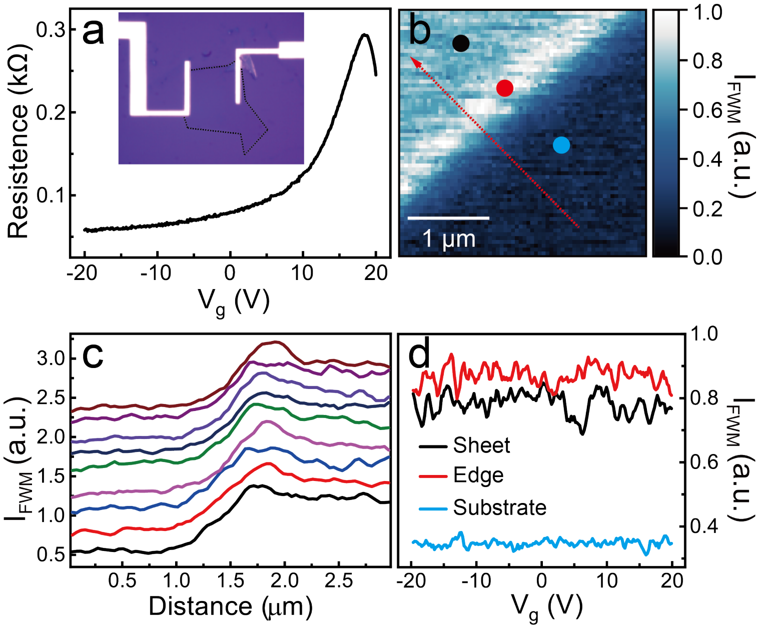

Figure S12. **FWM nano-imaging of graphene FET device.** (a) Graphene resistance as a function of back-gate voltage V_g. Inset of (a) shows the bright field optical image of the graphene FET device. (b) FWM imaging of graphene area near edge at V_g = -20 V. (c) Line traces along the red arrow in (b) of integrated FWM intensity gated at different Vg. Line traces from bottom to top correspond to V_g changed from -20 V to 20 V in steps of 5 V. (d) Back-gate voltage ramp and FWM at 3 locations marked in (b): on graphene sheet (black), on edge (red), and on substrate (blue).

Supplementary Note 13: AFM and nano-FWM image overlay.

From analysis and superposition of AFM and FWM images, it is evident that the near-edge FWM signal peaks at a distance close to the value of spatial delocalization inside from the actual graphene edge (Fig. S13). At the actual edge the signal decays to the nearly negligible substrate signal level. This behavior is in fact a further verification of the nonlocal Doppler broadening model which predicts that the nonlinear susceptibility decreases at near-field wave vectors larger than the cutoff value $q_{max} \sim \Delta\omega/v_F$. The signal cannot have a sharp peak at the edge as this would require a nonlinear polarization spectrum extending to larger wave vectors than this cutoff. We therefore expect the spatial distribution of the nonlinear polarization to be a smooth function extending over a scale of ~100's nm from its peak and smoothly dropping to zero at the edge as supported by the data. There is a difference in delocalization between graphene internal edge (white dashed line between folded and unfolded area), and edge-to-substrate. Both the folded area and unfolded area close to the internal edge show an enhanced FWM signal and a broader delocalization than the edge close to the substrate.

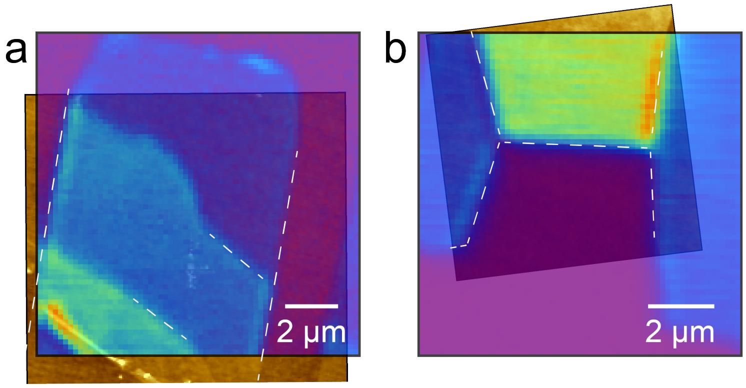

Figure S13. **AFM and nano-FWM image overlays of the graphene samples as shown in Fig. 1e (a) and Fig. 2b (b).** These overlays show that signal distribution near the outer edge is smooth, symmetric, and its peak offset from the graphene edge. Each AFM image is overlaid with the corresponding FWM image with 75% opacity. The dashed white lines indicate the actual graphene edges.

Supplementary Note 14: Coherent phonon effect on graphene FWM.

In principle, the original bandwidth of our ~10 fs pulses allows for coherent excitation of vibrational modes in graphene. However, with the particular spectral filtering applied in this experiment for maximizing weak FWM signals from nanoscopic sample volumes, the effective bandwidth is reduced to ~0.15 eV. This does not allow exciting the prevalent G-mode (0.196 eV) and puts the D-mode (0.167 eV) at the very far shoulder of the laser spectrum, where the corresponding FWM intensity is already close to the noise floor. On the other hand, low-frequency phonons such as shear modes and breathing modes that exist in few-layer graphene are in principle accessible within our spectral bandwidth [14]. However, its detection would either require long time delays beyond the pulse replica spacing accessible by spatial light modulator (SLM) pulse shaping, and/or long measurement times to achieve an appreciable signal, and thus prone to drift. The sensitivity of our technique to actually probe coherently excited vibrational modes can be improved in an implementation of ultrafast coherent anti-Stokes Raman spectroscopy [15] for non-resonant FWM background suppression.

Supplementary Note 15: Theoretical model of Doppler broadening

A theory of the third order nonlinear response of graphene including spatial dispersion (or space-nonlocal effects) is still lacking, and in fact our results provide a motivation for its future development. Nevertheless, and as outlined in the main manuscript, the effect of a small tip size (or large wavenumbers of surface-plasmon pump fields) on the third order nonlinear response of graphene can be predicted qualitatively with high degree of confidence, because it follows from the general hierarchy of the density matrix equations solved by the method of successive perturbations [16]. Indeed, the linear perturbation of the density matrix by a Fourier harmonic of the field with frequency ω_j and in-plane wave vector $\boldsymbol{q}_j$ is [17]

$$\rho^{(1)}_{mn} = \frac{V_{mn}(\boldsymbol{q}_j)(\rho^{(0)}_{nn} - \rho^{(0)}_{mm})}{\omega_j - \frac{E_m - E_n}{\hbar} + i\mathit{\Gamma}}, \tag{1}$$

where $V_{mn}(\boldsymbol{q}_j)$ are matrix elements of the interaction Hamiltonian and superscripts (0) denote unperturbed elements of the density matrix. Energies of the electron states $E_m(\boldsymbol{k_m})$ and $E_n(\boldsymbol{k_n})$ depend on quasimomenta that have to satisfy the momentum conservation $\boldsymbol{k}_m - \boldsymbol{k}_n = \boldsymbol{q}_j$. If the characteristic spatial scales of all fields are larger than the de Broglie wavelengths of carriers, the energy difference in Eq. (1) can be expanded in powers of $\boldsymbol{q}_j$ to give

$$\rho^{(1)}_{mn} = \frac{V_{mn}(\boldsymbol{q}_j)\left(\rho^{(0)}_{nn}(\boldsymbol{k}_n) - \rho^{(0)}_{mm}(\boldsymbol{k}_n) - \frac{\partial \rho^{(0)}_{mm}}{\partial \boldsymbol{k}}\boldsymbol{q}_j\right)}{\omega_j - \omega_{mn} + i\mathit{\Gamma} - \frac{1}{\hbar}\frac{\partial E_m}{\partial \boldsymbol{k}}\boldsymbol{q}_j}, \tag{2}$$

where $\omega_{mn} = \frac{E_m(\boldsymbol{k}_n) - E_n(\boldsymbol{k}_n)}{\hbar}$ is the transition frequency neglecting the change in electron momentum and the quantity $\frac{1}{\hbar}\frac{\partial E_m}{\partial \boldsymbol{k}}\boldsymbol{q}_j$ describes the Doppler shift of the optical frequency.

In higher orders of perturbation $\rho^{(\alpha)}_{mn}$ the frequencies ω_j and wave vectors $\boldsymbol{q}_j$ will be replaced by the "combination" frequencies $\omega^{(2)} = \omega_1 \pm \omega_2$, $\omega^{(3)} = \omega_1 \pm \omega_2 \pm \omega_3$ and wave vectors $\boldsymbol{q}^{(2)} = \boldsymbol{q}_1 \pm \boldsymbol{q}_2$, $\boldsymbol{q}^{(3)} = \boldsymbol{q}_1 \pm \boldsymbol{q}_2 \pm \boldsymbol{q}_3$, etc. The density matrix elements will then acquire resonant denominators which depend on the corresponding Doppler shifts:

$$\rho^{(\alpha)}_{mn} = \frac{F(\rho^{(\alpha-1)}_{pq}, \ldots, \rho^{(1)}_{pq})}{\omega^{(\alpha)} - \omega_{mn} + i\mathit{\Gamma} - \frac{1}{\hbar}\frac{\partial E_m}{\partial \boldsymbol{k}}\boldsymbol{q}^{(\alpha)}}, \tag{3}$$

where the numerator depends on the density matrix elements found in previous orders of perturbation. The measured average value of a given Fourier harmonics of the third-order nonlinear current density, say $\boldsymbol{j}^{(3)}(2\omega_1 - \omega_2, 2\boldsymbol{q}_1 - \boldsymbol{q}_2)$, can be found by taking the trace of the density matrix $\rho^{(3)}_{mn}(2\omega_1 - \omega_2)$ with the corresponding Fourier harmonic of the current density operator matrix $\boldsymbol{j}_{nm}(2\boldsymbol{q}_1 - \boldsymbol{q}_2)$, where

$$\boldsymbol{j}_{nm}(\boldsymbol{q}) = 2\langle n|e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}\hat{\boldsymbol{j}}|m\rangle$$

and $\hat{\boldsymbol{j}} = -ev_F\hat{\boldsymbol{\sigma}}$; see [17]. This step involves integration over electron quasimomenta.

After that, the inverse Fourier transform can be taken to calculate the third-order nonlinear current for a given frequency-wavevector distribution of the surface plasmon fields $\mathcal{E}_{1,2}(\omega, \boldsymbol{q})$, which include large wavenumbers up to $q_{max} \sim 2\pi/R$, where R is the tip radius. This procedure is conceptually straightforward but extensive, as evidenced by the amount of effort already needed to derive the *second-order* nonlinear current including the effects of spatial dispersion [17] and the third-order nonlinear current *neglecting* the effects of spatial dispersion [18]. However, the qualitative impact of the spatial dispersion of the nonlinear susceptibility on the observed nonlinear signal is possible to predict just from the general structure of the perturbative density matrix elements in Eq. (3). Indeed, when the spatial dispersion effects are included, the frequencies in resonant denominators in Eq. (3) acquire extra Doppler broadening factors $\sim v_F q^{(\alpha)}$, where we replaced $\frac{1}{\hbar}\frac{\partial E_m}{\partial \boldsymbol{k}}$ with the characteristic velocity v_F of electrons in graphene. This factor is most important in terms which originate from second-order corrections and contain the smallest frequencies,

$$\rho^{(2)}_{mn} \propto \frac{1}{\Delta\omega + i\Gamma - v_F q^{(2)}}, \tag{4}$$

where $\Delta\omega$ is either frequency difference $\omega_1 - \omega_2$ of two quasi-monochromatic pump fields or the frequency width of a broadband pump pulse. Even for a very broad wavevector spectrum of the near field at the tip apex, reaching maximum values of $q_{max} \sim 2\pi/R > \Delta\omega/v_F$, the contributions of the field spatial harmonics with $q > \Delta\omega/v_F$ to the nonlinear signal intensity will be suppressed as $1/q^2$. Therefore, the wavevector spectrum of the nonlinear signal will be determined by the harmonics with a maximum value of $q \sim \Delta\omega/v_F$. As a consequence, the tip-enhanced FWM signal shifts from and then decays to zero at the actual graphene edge since the nonlinear susceptibility decreases at larger near-field wave vectors. This will limit the spatial localization to scales of order $\Delta L \sim 2\pi v_F/\Delta\omega$. For the observed FWM signal bandwidth of $\hbar\Delta\omega \sim 40$ meV and $v_F = 10^6$ m/s we obtain $\Delta L \sim 100$ nm, in qualitative agreement with experiment. Note that the pump pulses have a broader spectrum than the FWM spectrum showed. This is necessarily narrowed by the required spectral filtering to a FWM of ~ 40 meV for imaging, which determines the length scale of FWM delocalization. One can see that the resulting loss in spatial resolution in FWM imaging of graphene is largely due to a high velocity v_F of all electrons. In other materials such as Drude metals or wide-gap semiconductors with largely parabolic electron dispersion, a significant or even dominant contribution to the nonlinear susceptibility may come from electrons with low group velocities, in which case the nonlocal Doppler broadening less prominent.

We emphasize that these are merely order-of-magnitude arguments that only qualitatively explain the mechanism of Doppler broadening on the near-field FWM response of graphene, which motivates a future quantitative derivation of the nonlocal third-order nonlinear response of graphene and other materials.

* Corresponding email: belyanin@physics.tamu.edu

† Corresponding email: markus.raschke@colorado.edu

[1] Cheng, J. L., Vermeulen, N. & Sipe, J. E. Third order optical nonlinearity of graphene. *New J. Phys.* **16**, 53014 (2014).

[2] Jiang, T. *et al.* Gate-tunable third-order nonlinear optical response of massless Dirac fermions in graphene. *Nat. Photon.* **12**, 430–436 (2018).

[3] Kravtsov, V. *et al.* Enhanced third-order optical nonlinearity driven by surface-plasmon field gradients. *Phys. Rev. Lett.* **120**, 203903 (2018).

[4] Babadjanyan, A. J., Margaryan, N. L. & Nerkararyan, K. V. Superfocusing of surface polaritons in the conical structure. *Journal of Applied Physics* **87**, 3785–3788 (2000).

[5] Stockman, M. I. Nanofocusing of optical energy in tapered plasmonic waveguides. *Phys. Rev. Lett.* **93**, 137404 (2004).

[6] Kravtsov, V., Ulbricht, R., Atkin, J. M. & Raschke, M. B. Plasmonic nanofocused four-wave mixing for femtosecond near-field imaging. *Nat. Nanotech.* **11**, 459–464 (2016).

[7] Groß, P. *et al.* Plasmonic nanofocusing–grey holes for light. *Advances in Physics: X* **1**, 297–330 (2016).

[8] Yao, X., Tokman, M. & Belyanin, A. Efficient nonlinear generation of thz plasmons in graphene and topological insulators. *Phys. Rev. Lett.* **112**, 055501 (2014).

[9] Constant, T. J., Hornett, S. M., Chang, D. E. & Hendry, E. All-optical generation of surface plasmons in graphene. *Nat. Phys.* **12**, 124127 (2016).

[10] Ju, L. *et al.* Graphene plasmonics for tunable terahertz metamaterials. *Nat. Nanotech.* **6**, 630–634 (2011).

[11] Fei, Z. *et al.* Gate-tuning of graphene plasmons revealed by infrared nano-imaging. *Nature* **487**, 82–85 (2012).

[12] Yan, H. *et al.* Tunable infrared plasmonic devices using graphene/insulator stacks. *Nat.*

Nanotech. **7**, 330–334 (2012).

[13] Chen, J. *et al.* Optical nano-imaging of gate-tunable graphene plasmons. *Nature* **487**, 77–81 (2012).

[14] Ishioka, K. *et al.* Ultrafast electron-phonon decoupling in graphite. *Phys. Rev. B* **77**, 121402 (2008).

[15] Oron, D., Dudovich, N. & Silberberg, Y. Femtosecond phase-and-polarization control for background-free coherent anti-stokes raman spectroscopy. *Phys. Rev. Lett.* **90**, 213902 (2003).

[16] Il'inskii, Y. & Keldysh, L. *Electromagnetic Response of Material Media* (Springer US, 1994).

[17] Wang, Y., Tokman, M. & Belyanin, A. Second-order nonlinear optical response of graphene. *Phys. Rev. B* **94**, 195442 (2016).

[18] Mikhailov, S. A. Quantum theory of the third-order nonlinear electrodynamic effects of graphene. *Phys. Rev. B* **93**, 085403 (2016).