nature photonics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41566-017-0054-7

In the format provided by the authors and unedited.

Broadband gate-tunable terahertz plasmons in graphene heterostructures

Baicheng Yao[1,2,7]*, Yuan Liu[3,4], Shu-Wei Huang[1,8], Chanyeol Choi[1], Zhenda Xie[1,9], Jaime Flor Flores[1], Yu Wu[2], Mingbin Yu[5,10], Dim-Lee Kwong[5,11], Yu Huang[3,4], Yunjiang Rao[2], Xiangfeng Duan[4,6]* and Chee Wei Wong[1]*

[1]Fang Lu Mesoscopic Optics and Quantum Electronics Laboratory, University of California, Los Angeles, CA, USA. [2]Key Laboratory of Optical Fiber Sensing and Communications (Education Ministry of China), University of Electronic Science and Technology of China, Chengdu, China. [3]Department of Materials Science and Engineering, University of California, Los Angeles, CA, USA. [4]California Nanosystems Institute, University of California, Los Angeles, CA, USA. [5]Institute of Microelectronics, Singapore, Singapore. [6]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA. Present addresses: [7]Cambridge Graphene Center, University of Cambridge, Cambridge, UK. [8]Department of Electrical, Computer, and Energy Engineering, University of Colorado Boulder, Boulder, CO, USA. [9]National Laboratory of Solid State Microstructures and School of Electronic Science and Engineering, Nanjing University, Nanjing, China. [10]Shanghai Institute of Microsystem and Information Technology, and Shanghai Industrial Technology Research Institute, Shanghai, China. [11]Institute for Infocomm Research, Singapore, Singapore. *e-mail: yaobaicheng@uestc.edu.cn; xduan@chem.ucla.edu; cheewei.wong@ucla.edu

Supplementary Information of

Broadband gate-tunable THz plasmons in graphene heterostructures

Baicheng Yao[1,2,a,*], Yuan Liu[3,4], Shu-Wei Huang[1,b], Chanyeol Choi[1], Zhenda Xie[1,c], Jaime Flor Flores[1], Yu Wu[2], Mingbin Yu[5,d], Dim-Lee Kwong[5,e], Yu Huang[3,4], Yunjiang Rao[2], Xiangfeng Duan[4,6*], and Chee Wei Wong[1*]

[1] Fang Lu Mesoscopic Optics and Quantum Electronics Laboratory, University of California, Los Angeles, CA 90095, United States

[2] Key Laboratory of Optical Fiber Sensing and Communications (Education Ministry of China), University of Electronic Science and Technology of China, Chengdu 610054, China

[3] Department of Materials Science and Engineering, University of California, Los Angeles, CA 90095, United States

[4] California Nanosystems Institute, University of California, Los Angeles, CA 90095, United States

[5] Institute of Microelectronics, Singapore 117685, Singapore

[6] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, United States

[a] current affiliation: Cambridge Graphene Center, University of Cambridge, CB3 0FA, United Kingdom

[b] current affiliation: University of Colorado Boulder, Boulder, CO 80309, United States

[c] current affiliation: Nanjing University, Jiangsu 210008, China

[d] current affiliation: Shanghai Institute of Microsystem and Information Technology, and Shanghai Industrial Technology Research Institute, Shanghai 200050, China

[e] current affiliation: Institute for Infocomm Research, Singapore 138632, Singapore

* Correspondence to: yaobaicheng@uestc.edu.cn; xduan@chem.ucla.edu; cheewei.wong@ucla.edu

S1. Comparison of THz optical sources

S2. Theoretical analysis

S3. Fabricating the gated dual-layer-graphene - nitride waveguide for THz plasmons

S1. Comparison of THz sources

Motivated by graphene's unique tunability, long-lived collective excitation and its extreme light confinement, we find an attractive potential of graphene plasmons to realize tunable THz sources [S1]. Table S1 compares some typical THz sources (0.1~50 THz) reported previously [S2-S11]. The unique advantage of the THz plasmon generation in graphene heterostructures is its wide tunability, approximately 10 times higher than the state-of-art tunable QCLs and undulators.

Table S1 | Comparison of the THz optical sources

Type	Output frequency (THz)	Tunability (THz)	Tuning frequency via	Potential to be fast	Ref.
Conventional QCLs	20	n/a			[S2-S3]
DFG based on crystals	1.4	≈ 4.7 (4.7 to 9.4 THz)	changing seed frequency	No	[S4]
Frequency comb based on microresonator	1.61	n/a			[S5]
Frequency comb + QCL	2.5	n/a			[S6]
QCL + DFG	34	≈ 0.74 (33.72 to 34.46 THz)	tuning temperature	No	[S7]
QCL + gratings	38.4	≈ 0.60 (38.07 to 38.67 THz)	tuning temperature	No	[S8]
QCL + DBR + DFB	3.8	≈ 0.6 (3.4 to 4 THz)	modulating static bias (temperature)	No	[S9]
Cherenkov DFG	3	≈ 3.6 (1.7 to 5.3 THz)	rotating the diffraction grating	No	[S10]
Helical undulator	0.1, 0.2, 0.4	not continuous	changing the diameter of nanowire	No	[S11]
Graphene plasmonic heterostructure	7	≈ 4.7 (4.7 to 9.4 THz) limited by EDFA	tuning gate voltage	Yes	This work

Table S2 compares graphene plasmon generation, observation and control in this work with the

state-of-literature techniques [S12-S20]. Figure S1 maps the performances of the state-of-art gate tunable graphene plasmons reported recently. Figure S1a shows, by using ultrathin Al_2O_3 dielectric barrier between dual layer graphene, we achieve an octave tunability, for the first time. Under single volt gating, Fermi level of the graphene atomic layers in our GSiNW can be modulated across the Dirac point. Figure S1b shows, the efficiency of gate tunability in this work is near 1 order higher than the published state-of-art works. Figure 1c maps that this work is unique using 'C+L' optical sources, which is cheap and widely applied in optical systems. Figure 1d highlights that compared to other graphene plasmon generations based optical nonlinearities, the on-chip waveguide design with ≈ 1 μm^2 mode field area enables this work without OPA or femtosecond pump, we apply a ps pulsed pump with 40 mW maximum average power (200 W peak power), the on-chip peak power density can reach 10 GW/cm^2, which is ≈ 1 order higher than previous reports based on out-of-plane implement. The GSiNW design enables the nonlinear process higher efficiency.

Table S2 | Comparison of the graphene plasmon generation and control.

Device design	Excitation		Observation method	Gate tunability	Ref.
	Scheme	Wavelength			
Monolayer graphene sample	Out-of-plane	Mid-infrared (9.7 to 11.2 μm) pump	*s*-SNOM	Yes	[S12-S13]
Monolayer graphene encapsulated with *h*-BN sample	Out-of-plane	Mid-infrared (ultrafast; ≈ 200fs) probe – infrared pump	*s*-SNOM	Yes	[S14]
Monolayer graphene nano-antenna sample	Out-of-plane	Mid-infrared (10.2 to 11.1 μm) illumination	*s*-SNOM	Yes	[S16]
Monolayer graphene nanoresonator sample	Out-of-plane	Mid-infrared (10 to 12) μm pump	*s*-SNOM	No	[S18]
Monolayer graphene nano-island sample	Out-of-plane SHG/THG	Mid-infrared source	Mid-infrared optical spectroscopy	Potential	[S17]
Monolayer graphene sample	Out-of-plane DFG	All-visible wavelength 1-kHz 100-fs mJ pump-probe	mJ pump-probe	No	[S20]
Monolayer graphene nanoribbon sample	Out-of-plane	Fourier transform infrared broadband source	Fourier transform spectroscopy	Yes	[S15]
Monolayer graphene nanoribbon sample	Out-of-plane	Broadband infrared source	Fourier transform spectroscopy	No	[S19]
Dual-layer heterogeneous graphene; integrated with on-chip waveguides	In-plane counter-pumped DFG	All near-infrared pulsed pump (C band) + CW probe (L band)	All near-infrared optical spectroscopy	Yes	this work

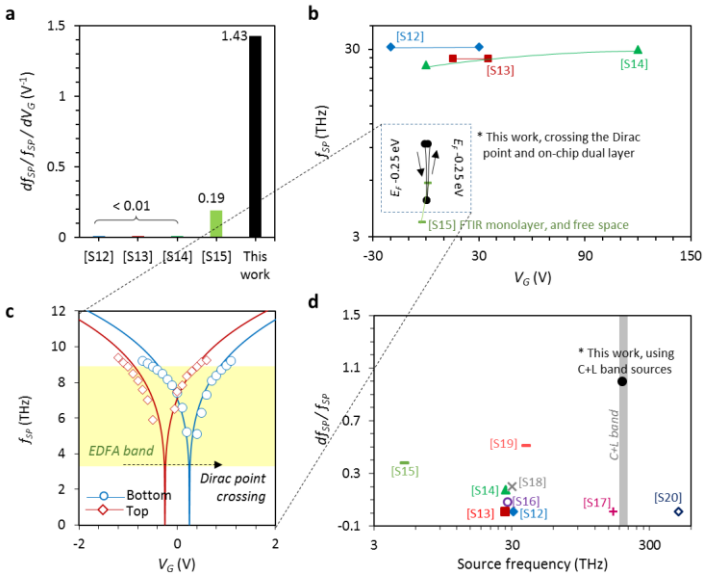

Figure S1 | Comparison on the gate-tunable THz plasmons. **a,** Comparison of gate-tuning efficiency [df_{sp}/f_{sp}/dV_G] for different studies. **b,** Comparison map of f_{SP} with V_G for different studies. **c,** Zoom-in of our f_{SP} versus V_G for the top and bottom graphene layers. **d,** Comparison map of [df_{sp}/f_{sp}] versus source frequency across a number of studies.

S2. Theoretical analysis

S2.1 Dual layer graphene – optical waveguide interaction

Figures S2.1a and S2.1b show the cross-sectional views of the dual-layer graphene nitride and the original nitride waveguide. Figure S2.1c shows the computed effective index dispersion of the TM fundamental mode in the graphene based silicon nitride waveguide, calculated via finite-element method with COMSOL commercial software. Here the index of silicon nitride material ranges from 1.9886 to 1.9904 (188 THz to 200 THz) [S21], the index of SiO_2 cladding is fixed at 1.4462, and the index of air is 1. To model the phase matching in dual-layer graphene structure, the effective graphene and pump-signal mode indices need to be examined. For the graphene-nitride structure, the fields of the TM fundamental mode transmitting along a conventional waveguide [S22] can be written as

$$B_x = \begin{cases} B_1 e^{ik_z t}\cos\left(\frac{k_2 h_{core}}{2}-\varphi\right)\exp k_1\left(\frac{h_{core}}{2}-y\right), y > \frac{h_{core}}{2} \\ B_2 e^{ik_z t}\cos(k_2 y-\varphi), -\frac{h_{core}}{2} < y < \frac{h_{core}}{2} \\ B_3 e^{ik_z t}\cos\left(\frac{k_2 h_{core}}{2}+\varphi\right)\exp k_3\left(\frac{h_{core}}{2}+y\right), y < \frac{h_{core}}{2} \end{cases} \quad \text{(S1)}$$

$$E_y = \begin{cases} \frac{ick_1 B_1}{\omega\epsilon_1} e^{ik_z t}\cos\left(\frac{k_2 h_{core}}{2}-\varphi\right)\exp k_1\left(\frac{h_{core}}{2}-y\right), y > \frac{h_{core}}{2} \\ \frac{ick_2 B_2}{\omega\epsilon_2} e^{ik_z t}\cos(k_2 y-\varphi), -\frac{h_{core}}{2} < y < \frac{h_{core}}{2} \\ -\frac{ick_3 B_3}{\omega\epsilon_3} e^{ik_z t}\cos\left(\frac{k_2 h_{core}}{2}+\varphi\right)\exp k_3\left(\frac{h_{core}}{2}+y\right), y < \frac{h_{core}}{2} \end{cases} \quad \text{(S2)}$$

$B_{1,2,3}$ are the maximum magnetic field intensities in air, core and SiO_2 under the core, and $k_2^2+k_z^2 = \epsilon_2(\omega/c)^2$, $-k_1^2+k_z^2 = \epsilon_1(\omega/c)^2$, $-k_3^2+k_z^2 = \epsilon_3(\omega/c)^2$ are the propagation constants. ϵ_1, ϵ_2, ϵ_3 are the permittivities of the zone 1, 2, and 3, respectively. Here $\epsilon_{1,2,3}=n_{1,2,3}^2$, φ is the phase constant, and $\omega=2\pi f$ is the frequency. For the propagation mode, the propagation constant k_y satisfies

$$tan(k_y h_{core}) = \left(\frac{k_y}{\epsilon_2}\frac{k_1}{\epsilon_1}+\frac{k_y}{\epsilon_2}\frac{k_3}{\epsilon_3}\right)\left(\frac{k_y}{\epsilon_2}\frac{k_y}{\epsilon_2}-\frac{k_1}{\epsilon_1}\frac{k_3}{\epsilon_3}\right) \quad \text{(S3)}$$

$$tan(2\varphi) = \left(\frac{k_y}{\epsilon_2}\frac{k_3}{\epsilon_3}-\frac{k_y}{\epsilon_2}\frac{k_1}{\epsilon_1}\right)\left(\frac{k_y}{\epsilon_2}\frac{k_y}{\epsilon_2}+\frac{k_1}{\epsilon_1}\frac{k_3}{\epsilon_3}\right) \quad \text{(S4)}$$

Because graphene is of considerable index n_g and conductivity σ_g, it can dramatically modify the boundary conditions. Referring the electromagnetic boundary conditions on the dual layer graphene layer

$$\epsilon_1 E_1 - \epsilon_2 E_2 = \rho_g, B_1 - B_2 = \sigma_g E_2 \quad \text{(S5)}$$

Here $\rho_g > 0$ and $\sigma_g > 0$ are the surface charge and conductivity of the layer. For $z \to \infty$, the simulated E-field distributions for the GSiNW are shown in Figure S2.1d. Light interacts with graphene layers via the evanescent field. Here graphene-Al_2O_3-graphene layer is of 0.4 nm + 30 nm + 0.4 nm thickness.

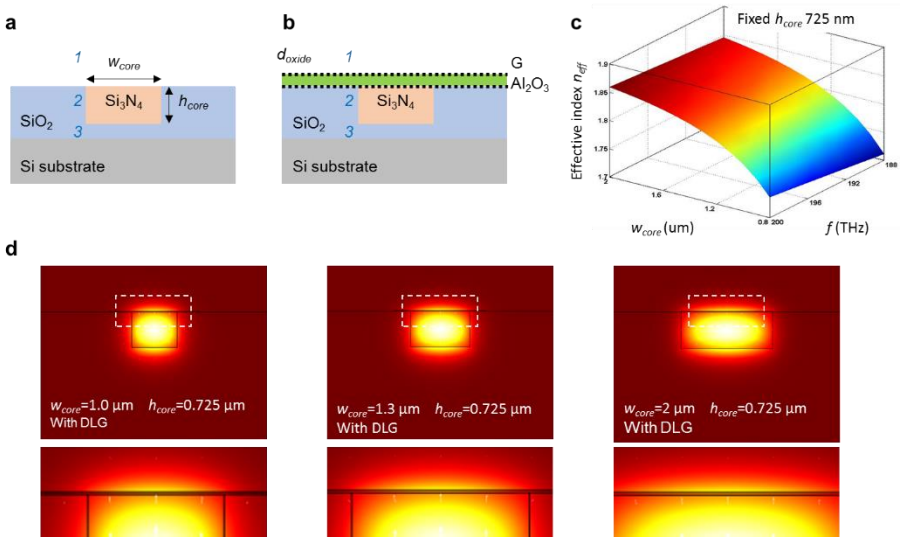

Figure S2.1 | Mode distributions. a, Cross-section of the original silicon nitride waveguide without graphene coverage. **b,** Cross-section of the silicon nitride waveguide with graphene-Al_2O_3-graphene coverage. **c,** Effective index dispersion of the fundamental TM mode in the silicon nitride waveguide, meshed in the waveguide width (w_{core}) and guided frequency (f) map. **d,** Simulated E-field distributions of the fundamental TM modes in the GSiNW with w_{core} = 1 μm, 1.3 μm and 2 μm. Here the graphene layers are assumed with a $|E_F|$ = 0.1 eV. DLG: dual-layer graphene.

S2.2 Phase matching in the DFG based plasmon generation

In the DFG based plasmon generation, energy converts from a pump photon (f_p in C band) to a signal photon (f_s in C band) and a plasmon (f_{SP} in THz band). During this process, momentum is conserved. Thus we write the energy matching and phase matching condition as

$$hf_s + hf_{SP} = hf_p, \overrightarrow{k_s} + \overrightarrow{k_{SP}} = \overrightarrow{k_p} \tag{S6}$$

Here h is the Planck constant, k_s, k_{sp} and k_p are the wavevectors of the signal, plasmon and pump. In optics, $k = 2\pi/\lambda = 2\pi n_{eff}/cT = 2\pi f n_{eff}/c$, where n_{eff} is the effective index and c is the light speed in vacuum. With the counter-propagation pump-signal geometry, we rewrite Eq. (S6) to be

$$\begin{cases} f_s n_s - f_{SP} n_{SP} = -f_p n_p \\ f_s + f_{SP} = f_p \end{cases} \tag{S7}$$

Here n_p, n_s, and n_{SP} are the effective indexes of the pump light, signal light, and the plasmon respectively. To satisfy the phase matching, f_p, f_s and n_{SP} should be selected and adjusted carefully:

$$\frac{f_s}{f_p} = \frac{n_{SP} - n_p}{n_{SP} + n_s}, n_{SP} > n_p \tag{S8}$$

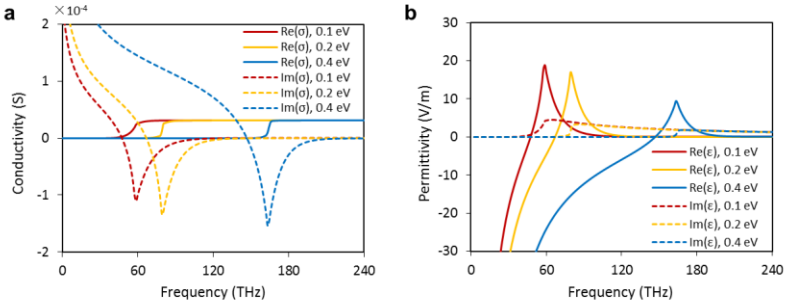

Figure S2.2.1 | Graphene conductivity and permittivity. a, Calculated conductivity of graphene, with Fermi level at 0.1 eV (red), 0.2 eV (yellow), and 0.4 eV (blue). **b,** Calculated permittivity of graphene, with Fermi level at 0.1 eV (red), 0.2 eV (yellow), and 0.4 eV (blue). Here the solid curves show the real parts while dashed curves show the imaginary parts.

In our measurements of the main text f_p is fixed at 195.8 THz (1531.9 nm) (in Supplementary Section S4.5, the pump wavelength is varied). The effective index n_p of the silicon nitride waveguide at f_p is ≈ 1.77 (waveguide with w_{core} = 1.3 μm and h_{core} = 0.75 μm). In our measurements, f_s is scanned from 192.3 THz to 177.5 THz (1560 nm to 1690 nm); the effective index n_s ranges from ≈ 1.77 to 1.75. The material index of graphene $n_g = n_{g,r} + in_{g,i}$ plays the key role in this equation. One can derive n_g from σ_g, as

$$\sigma_g(f, E_F, \tau, T) = \frac{ie^2(2\pi f - i/\tau)}{\pi\hbar^2}\left\{\frac{1}{\left(2\pi f+\frac{i}{\tau}\right)^2}\int_0^\infty \varepsilon\left[\frac{\partial f_d(\epsilon)}{\partial\epsilon} - \frac{\partial f_d(-\epsilon)}{\partial\epsilon}\right]d\epsilon - \int_0^\infty \left[\frac{f_d(-\epsilon)-f_d(\epsilon)}{(2\pi f+i/\tau)^2-4(\epsilon/\hbar)^2}\right]d\epsilon\right\} \quad \text{(S9)}$$

Specifically,

$$\sigma_{g,intra} = \frac{ie^2E_F}{\pi\hbar(2\pi f+\frac{i}{\tau})} \quad \text{(S10)}$$

$$\sigma_{g,inter} = \frac{ie^2E_F}{4\pi\hbar} ln\left[\frac{2|E_F|-\hbar(2\pi f+\frac{i}{\tau})}{2|E_F|+\hbar(2\pi f+\frac{i}{\tau})}\right] \quad \text{(S11)}$$

Hence,

$$\epsilon_g = \frac{-\sigma_{g,i}+i\sigma_{g,r}}{2\pi f\Delta} \quad \text{(S12)}$$

$$\left(n_{g,r} + in_{g,i}\right)^2 = \epsilon_{g,r} + i\epsilon_{g,i} \quad \text{(S13)}$$

$$n_{g,r} = sqrt\left(\frac{-\epsilon_{g,r}+\sqrt{{\epsilon_{g,r}}^2-{\epsilon_{g,i}}^2}}{2}\right), n_{g,i} = sqrt\left(\frac{-\epsilon_{g,r}+\sqrt{{\epsilon_{g,r}}^2-{\epsilon_{g,i}}^2}}{2}\right) \quad \text{(S14)}$$

In above equations, E_F is the Fermi level, $\tau = 10^{-13}$ s is the relaxation lifetime, T is the temperature, $f_d(\epsilon) = \{exp[(\epsilon-E_F)/k_BT]+1\}^{-1}$ is the Fermi-Dirac distribution, $\hbar = 1.05\times10^{-34}$ eV·s is the reduced Planck constant, $k_B = 1.3806505\times10^{-23}$J/K is the Boltzmann's constant, and $e = -1.6\times10^{-19}C$ is the unit charge. When graphene is gated, n_g is much higher than n_p or n_s [S23], corresponding to the f_{sp} much smaller than f_p or f_s. Figure S2.2.1a and Fig. S2.2.1b shows the calculated conductivity and the permittivity of graphene using the Kubo formalism [S24-S26]. When $\sigma_{g,i} > 0$, graphene can support surface plasmons.

When the plasmon frequency is lower than Landau damping regime, we get the momentum-frequency (k_{SP}-f) dispersion of graphene. With the boundary conditions, it could be approximately simplified as a quadratic function (see also Eq. S28-S29) as

$$k_{SP} = Af^2 \quad \text{(S15)}$$

$$A = \frac{(1+{n_{SP}}^2)h\pi^2}{2\alpha c v_F} \quad \text{(S16)}$$

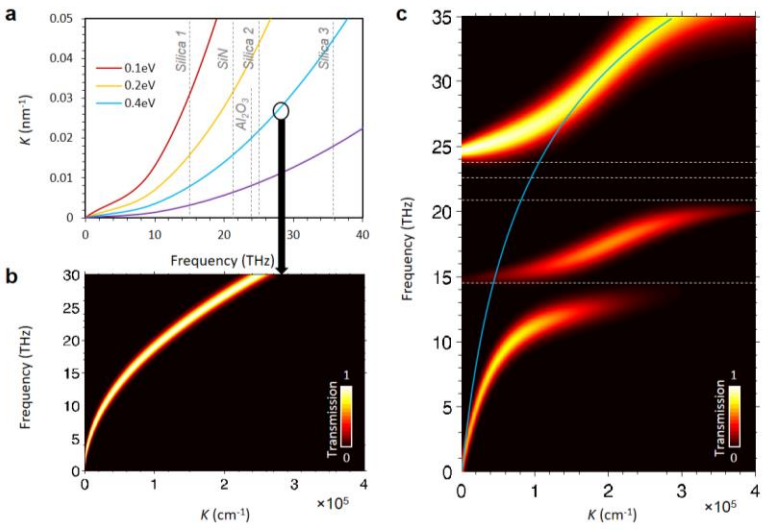

Figure S2.2.2 | Phase matching. a, Graphene *f*-***k*** dispersion, under Fermi level of 0.1 eV, 0.2 eV, 0.4 eV and 0.8 eV. **b,** Calculated $1/L_{SP}(\boldsymbol{k},f)$ of graphene with Fermi level 0.4 eV, based on RPA method. **c,** $1/L_{SP}(\boldsymbol{k},f)$ map describing the plasmon-phonon couplings in the GSiNW. Phonon frequencies are marked out with the dashed lines. In **b** and **c**, the values of the transmission ~ $1/L_{SP}(\boldsymbol{k},f)$ is normalized to be 1.

Here α is the fine structure constant, and E_F is the Fermi level of graphene. With the relationship of n_g and E_F, the calculated k_{SP}-f curves of graphene with Fermi level ranging from 0.1 eV to 0.8 eV are shown in Figure S1.2.2a. In Figure S1.2.2a, the grey lines show the phonon resonance locations of the Si_3N_4 and SiO_2. $f_{silica,1}$ = 14.55 THz, $f_{silica,2}$ = 24.18 THz, $f_{silica,3}$ = 36.87 THz, f_{SiN} = 21.89 THz, and f_{Al2O3} = 22.4 THz [S27-S33].

Furthermore, by using random phase approximation (RPA) method [S30-S33], we calculate the plasmon coupling based loss $L_{SP}(\boldsymbol{k},f)$ along the GSiNW, with consideration of the phonon couplings.

$$L_{SP}(\boldsymbol{k},f) = -Im\left\{1 - \frac{e^2}{2\boldsymbol{k}\epsilon_1}\prod_0(\boldsymbol{k},f) - \sum_j f_{ph,j}\prod_0(\boldsymbol{k},f)\right\} \quad \text{(S17)}$$

$$\prod_0(\boldsymbol{k},f) = -\frac{g_s}{4\pi^2}\sum\int \frac{f_d(\epsilon_s) - f_d(\epsilon_{s\boldsymbol{k}})}{2\pi f\hbar + \frac{i\hbar}{\tau} + \epsilon_s - \epsilon_{s\boldsymbol{k}}} d\boldsymbol{k}F(s,\boldsymbol{k}) \quad \text{(S18)}$$

Here $f_{ph,j}$ is the phonon resonances, g_s = 4, $f_d(\epsilon)$ the Fermi-Dirac distribution, $\epsilon_s = sv_F$, $\epsilon_{s\boldsymbol{k}} = sv_F\boldsymbol{k}$, $s = \pm 1$, $F(s,\boldsymbol{k})$ is the band overlap function of Dirac spectrum, which equals 1 for the waveguide geometry. Figure S2.2.2b provides the simulated $1/L_{SP}(\boldsymbol{k},f)$ map of graphene for the Fermi level at 0.4 eV, without considering the plasmon-phonon couplings. Figure S2.2.2c shows the RPA map with consideration of the phonon couplings.

S2.3 Plasmon enhanced 2nd-order nonlinearity and the THz frequency generation

Graphene is a single atomic layer with honeycomb structure, therefore, second-order nonlinear effects are described by the second-order surface nonlinear conductivity [S34]. With light transmitting along

graphene with a wavevector $\boldsymbol{k}$ parallel to the 2D layer plane, the second-order nonlinear polarizability $\chi^{(2)}$ can be large [S35]. In graphene, an effective $\chi^{(2)}$ can be written as

$$\frac{\partial^2 \chi_{ijk}^{(2)}}{\partial \boldsymbol{k}^2} = \frac{e^2}{4\pi^4\hbar^2 f_p f_s}\left\{\left[\frac{f(\boldsymbol{k_1})-f(\boldsymbol{k_3})}{\omega_{31}-2\pi f_p - i\gamma}+\frac{f(\boldsymbol{k_1})-f(\boldsymbol{k_2})}{2\pi f_s-\omega_{21}-i\gamma}\right]\frac{\mu_{32}^i v_{31}^i v_{21}^k}{\omega_{32}-2\pi f-i\gamma}-\left[\frac{f(\boldsymbol{k_1})-f(\boldsymbol{k_3})}{\omega_{31}-2\pi f_p-i\gamma}+\frac{f(\boldsymbol{k_2})-f(\boldsymbol{k_3})}{2\pi f_s-\omega_{23}-i\gamma}\right]\frac{\mu_{21}^i v_{31}^i v_{32}^k}{\omega_{21}-2\pi f-i\gamma}\right\} \quad \text{(S19)}$$

$$\mu_{ab} = \frac{iev_F}{\omega_{ab}} < a|\boldsymbol{\sigma_g}|b >, v_{ab} = v_F < a|\boldsymbol{\sigma_g}|b >, \omega_{ab} = \frac{E(\boldsymbol{k_a})-E(\boldsymbol{k_b})}{\hbar} \quad \text{(S20)}$$

Here $f(\boldsymbol{k})$ is the occupation number state $\boldsymbol{k}$, $\boldsymbol{k_1}$, $\boldsymbol{k_1}$ and $\boldsymbol{k_3}$ satisfy $\boldsymbol{k_1}+\boldsymbol{k_p}=\boldsymbol{k_3}$, $\boldsymbol{k_1}+\boldsymbol{k_s}=\boldsymbol{k_2}$. Here $\boldsymbol{\sigma_g}$ is the 2D Pauli matrix vector, $<a>$ and $<b>$ are the states, γ is the scattering rate, and $v_F=E_F/(\hbar\boldsymbol{k})$ is the Fermi velocity. By approximating $k_BT \rightarrow 0$, $2\pi f >> v_F k$, $f_p \approx f_s$, along the graphene, the second-order nonlinear polarizability $\chi_{ijk}^{(2)}$ can be simplified as

$$\chi_{eff}^{(2)} = \frac{e^3}{4\pi^2\hbar^2}\frac{1}{k\sqrt{f_s f_p}}\left[\frac{\pi}{2}+\arctan\left(\frac{2\pi\sqrt{f_s f_p}-2v_F k_F}{\gamma}\right)\right] \quad \text{(S21)}$$

Here $\hbar k_F = \hbar(2m_e E_F)^{1/2}$ is the Fermi momentum. The simulated $\chi_{eff}^{(2)}$ is shown in Figure S2.3a: A higher E_F brings a lower $\chi_{eff}^{(2)}$. Here f_p is fixed as 1.93 THz. Hence we write the E-field intensity of the generated plasmon as

$$E_{SP} = \frac{\chi_{eff}^{(2)} E_{p,y}\left(\frac{h_{core}}{2}\right)E_{s,y}\left(\frac{h_{core}}{2}\right)}{L_{SP}} \quad \text{(S22)}$$

Here L_{SP} is shown in Eq. (S15). Referring to Eq. (S1) and Eq. (S2), here $E_{p,y}$ $E_{s,y}$ are the E-field of the pump and the signal respectively, with $z = ct/n_g$, $k_p = 2\pi f_p n_{eff,p}/c$, $k_s = 2\pi f_s n_{eff,s}/c$. The real part of Eq. (S22) can be approximately simplified as

$$E_{SP}(t) = \frac{1}{2}A_{SP}(t)cos\left[2\pi\frac{(f_p n_p + f_s n_s)}{n_{SP}}t\right]+\frac{1}{2}A_{SP}(t)cos\left[2\pi\frac{(f_p n_p - f_s n_s)}{n_{SP}}t\right] \quad \text{(S23)}$$

$$A_{SP}(t) = \frac{A_p(t)A_s(t)\chi_{eff}^{(2)}}{L_{SP}}\exp(-n_{g,i}) \quad \text{(S24)}$$

$A_{SP}(t)$ determines the loss of the surface plasmon wave. Here $A_p(t)$ and $A_s(t)$ are the amplitudes of the pump and signal respectively. In this equation, we get the frequency of the surface plasmon $f_{SP} = (f_p n_p + f_s n_s)/n_{SP}$, and the frequency of heterodyne beat $f_B = (f_p n_p + f_s n_s)/n_{SP}$. Referring to the DFG energy balance $f_{SP} = f_p - f_s$, the effective index of the plasmon n_{SP} satisfies

$$n_{SP} = \frac{f_p n_p + f_s n_s}{f_p - f_s} \quad \text{(S25)}$$

This equation corresponds to the Eq. (S8) perfectly. For f_p and f_s located in '*C+L*' optical communications band and with $f_{SP} \approx 8$ THz, the n_{SP} satisfying the phase matching condition could be approximately calculated to be ≈ 80. n_{SP} is also determined by f_{SP} and the Fermi level E_F, from Eq. (S15). Figure S2.3b

plots the graphene dispersion n_{SP} (f_p, f_s) at E_F = 0.1 eV, and the DFG phase matching n_{SP} from Eq. (S25) together. This figure shows that the DFG based graphene plasmon generation is related to E_F, f_p, f_s and the waveguide structure, concisely together in one figure.

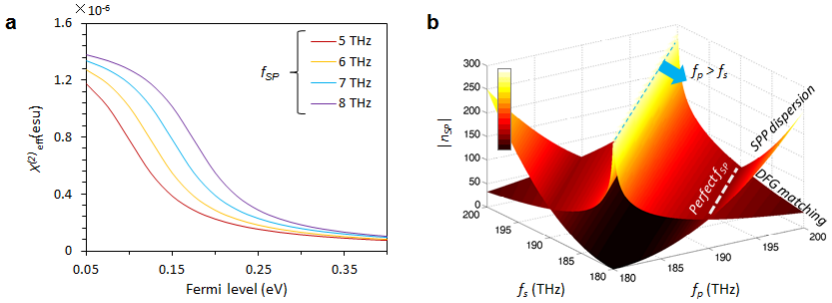

Figure S2.3 | 2nd-order nonlinear polarizability and effective indices of the surface plasmons. a, Calculated curves of $\chi_{eff}^{(2)}$ with f_{SP} at 5 THz (red), 6 THz (yellow), 7 THz (blue), and 8 THz (purple). **b,** For w_{core} = 1 μm and h_{core} = 725 nm, to satisfy the phase matching, n_{SP} is determined by both f_p and fs.

S2.4 Dual-layer graphene plasmon: coupling and gating

In Sections S2.2 and S2.3, phase matching of DFG based on monolayer graphene is analyzed, without considering the possible plasmon coupling of the separated graphene layers in the grphene-Al_2O_3-graphene system. When the distance between the two graphene layers is small enough, the graphene-Al_2O_3-graphene could be regarded as a topological insulator-like system in which plasmonic mode coupling can occur [S35-S39]. We schematically show the dual-layer graphene structure in Figure S2.4a. Here the thickness of Al_2O_3 is taken into consideration as d. Compared to single layer graphene, the situation in the gated graphene-Al_2O_3-graphene structure could be regarded as a capacitor: when stable, the top layer graphene charges $+Q$, and the bottom layer graphene charges $-Q$. The Fermi level of a monolayer graphene is written as [S40]

$$E_F = \hbar|v_F|\sqrt{\pi n} \tag{S26}$$

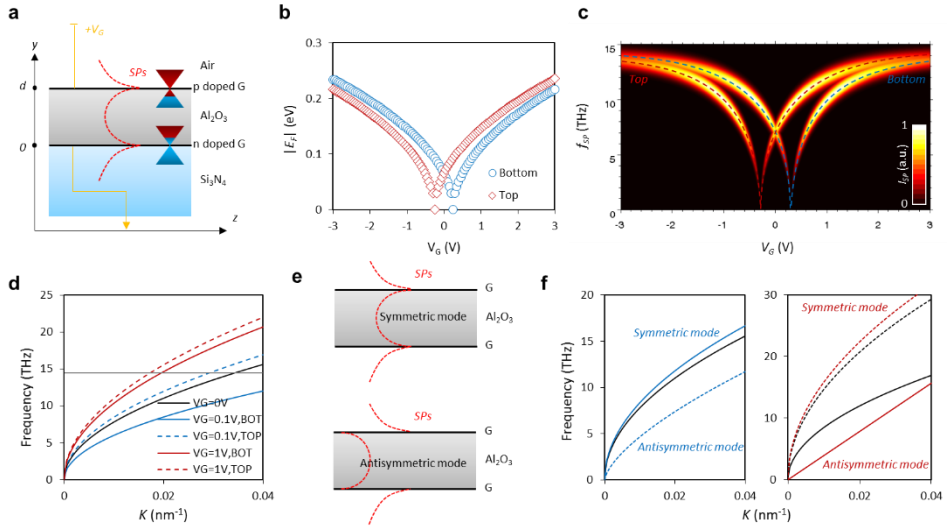

Figure S2.4 | Graphene-Al_2O_3-graphene system. a, Schematic configuration. **b,** Correlation of gate voltage and Fermi level. Blue circles denote the bottom layer and red diamonds denote the top layer. **c,** Simulated plasmon dispersions on the top and bottom graphene layers. When V_G = 0 V, the two graphene layers have the same E_F of ≈ 50 meV intrinsically. **d,** Dispersions under different V_G, without interlayer coupling. Solid curves denote the independent bottom layer and dashed curves denote the independent top layer. **e,** Symmetric and antisymmetric modes. **f,** Dispersions of coupled modes, which are mode-split from the original ones. Black curves: independent modes (solid: bottom layer; dashed: top layer). Left panel denotes V_G = 0V and right panel denotes V_G = -1V.

We note that the initial Fermi levels of the top and the bottom layer graphene would be different (E_T, E_B): once the graphene-Al_2O_3-graphene capacitor formed, the top layer graphene is positively charged while the bottom layer is negatively charged, $n_{electron}=n_{hole}=Q/eS_g=C_GV_G/eS_g$. That means, the carrier densities of the top layer (n_T) graphene and the bottom layer graphene (n_B) would be different. Considering CVD graphene in air is *p*-doped initially, $E_T > E_B$ when $V_G > 0$. Assuming the top and bottom layer graphene has the same size S_g = 80×20 μm^2, a capacitance C_G=2×10^{-7} F/cm^2, and the initial Fermi levels (before gating) E_{T0} = E_{B0} = - 50 meV (n_{hole_0} ~ 2×10^{11}/cm^2), Figure S2.4b shows the resulting computed V_G-E_F correlation. Since E_F determines the dispersion of graphene plasmon, Figure S2.4b predicts that two plasmons with different f_{SP} could be generated simultaneously and tuned differently with gate voltage in the graphene-Al_2O_3-graphene system. Figure S2.4c simulates the dispersions of the dual-layer graphene, without interlayer coupling. Figure S2.4d simulates the dispersions of top and bottom graphene surface plasmons, with reference to the initial one, E_{T0} = E_{B0} = - 50 meV at V_G = 0 V.

However, when the interlayer coupling distance *d* is small enough, the two independent plasmon modes would couple with each other to form two hybrid modes, symmetric and antisymmetric. In the low frequency regime, the dispersions of the symmetric and antisymmetric modes are described as:

$$f_{sym} = \frac{1}{2\pi}\left[\frac{2e^2}{\varepsilon}(E_T + E_B)k_{sp}\right]^{1/2} \quad \text{(S28)}$$

$$f_{asym} = \frac{1}{2\pi}\left[\frac{4e^2 E_T E_B d}{\varepsilon(E_T+E_B)}\right]^{1/2} k_{sp} \quad \text{(S29)}$$

Here ε is the background permittivity and d is the dielectric layer thickness. In Figures S2.4e and S2.4f, we show the calculated dispersions of f_{op} (k_{sp}) and f_{ac}(k_{sp}), with $e^2/\varepsilon \approx 7\times10^5$ THz^{2}nm·eV^{-1}, and d = 30 nm. Plasmon coupling further splits the dispersion curves in THz region.

S2.5 Considerations of DFG versus FWM

Graphene also has large $\chi^{(3)}$, which offers third-order optical nonlinearity, e.g. four wave mixing (FWM) [S41]. One might wonder if the enhancement of the signal is plausible from FWM instead of DFG [S42-S45]. Here analysis is shown theoretically to exclude the influence of FWM, in our pump-signal counter-launched configuration. In a typical degenerate FWM process, the photon energy transfers from pump to signal and idler, with energy and momentum matching. When the propagation directions of the pump and the signal are opposite, once FWM occurs we have

$$2f_p = f_s + f_i \quad \text{(S30)}$$

$$2\vec{k}_p = -\vec{k}_s + \vec{k}_i \quad \text{(S31)}$$

Here f_p, f_s and f_i are the frequencies of the pump, signal and idler, $\hbar k_p$, $\hbar k_s$ and $\hbar k_i$ are the momentums of the pump, signal and idler, $k=2\pi f n_{eff}/c$, respectively. The dispersion could be written as

$$2f_p(n_{eff,i} - n_{eff,p}) = f_s(n_{eff,i} + n_{eff,s}) \quad \text{(S32)}$$

Here $n_{eff,i}$, $n_{eff,p}$ and $n_{eff,s}$ are the effective mode indexes. To satisfy this equation when the frequency difference of f_p and f_s is smaller than 5 THz, $n_{eff,i}$ would have to be ≈ 3 times larger than $n_{eff,p}$ or $n_{eff,s}$. However, for the FWM-generated mode, its $n_{eff,i}$ cannot be larger than the index of the waveguide core. Hence FWM cannot occur in our counter-launched pump-signal configuration in this case.

S3. Fabricating the gated dual-layer-graphene - nitride waveguide for THz plasmons

S3.1 Fabrication process flow

Figure S3.1 shows the fabrication process steps of graphene on the silicon nitride waveguides (GSiNWs). As shown in *step 1*, the chips are fabricated at the Institute of Microelectronics Singapore, with the silicon nitride waveguide buried in SiO_2 cladding. There are 4 straight waveguides in every chip with a width of 1 μm, and length of ≈ 3 mm. The undercladding oxide is 3 μm thick, the height of the waveguide core is 725 nm, and the top oxide cladding is 2.5 μm. The chip is chemically etched by using wet buffered oxide etching (BOE) method in *step 2* (plasma-based dry etching is also available). After etching, the distance between the core and the top oxide surface is less than 20 nm, ensuring the strong light-graphene interaction. In *step 3*, a chemical vapor deposition (CVD) grown monolayer graphene is transferred onto

the chip using wet transfer, followed by photolithography patterning and oxygen plasma etching. This graphene layer serves as the bottom layer graphene with a size of 100 μm× 40 μm. Next, the Ti/Au (20/50 nm) contact pad is deposited using electron beam evaporation, working as source-drain electrodes. By using the source and drain, resistance of the bottom layer graphene could be measured. In *step 5*, a thin 30 nm layer of Al_2O_3 in deposited using atomic layer deposition (ALD), providing sufficient capacitance for the graphene based semiconductor chip. Finally, as shown in *step 6*, on the top of the Al_2O_3 insulator, another graphene layer is transferred, aligned and linked with the gate.

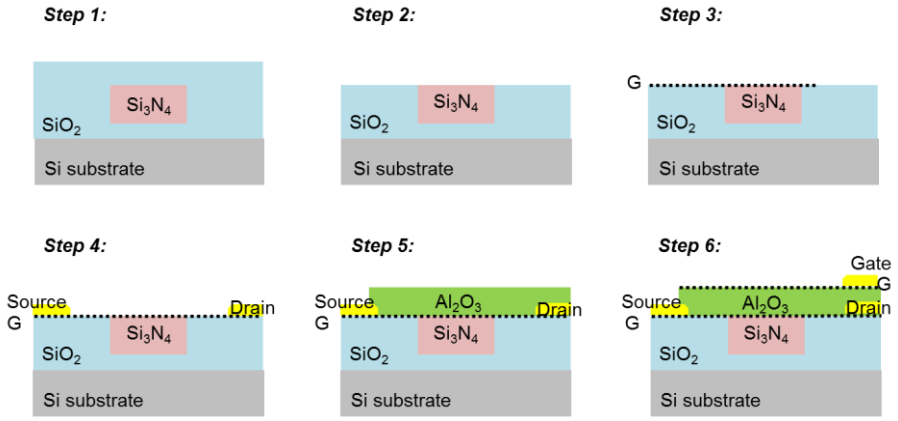

Figure S3.1 | Nanofabrication process of the dual-layer-graphene nitride plasmon structure. The graphene layers are transferred onto the nitride waveguide along with the source-drain-gate electrodes and the Al_2O_3 dielectric barrier layer deposition.

S3.2 Tuning the graphene – nitride DFG-plasmon interactions in the dual-layer graphene structures

Figure S3.2a (left) illustrates a top-view optical micrograph of the etched silicon nitride waveguides (SiNWs). The edge of the etched and the non-etched areas is clear. To reduce the scattering and coupling loss of the etched waveguides, the inverse taper couplers at the input and output facets are carefully protected by photoresist. Figure S3.2a (right) shows the after-etch oxide thickness at random locations, with an uncertainty of ± 10 nm. The oxide thickness refers the distance between the top surface and the bottom Si substrate of a chip. The thickness data is measured by using an optical interferometer at 480 nm wavelength, with the SiO_2 refractive index fixed at 1.4594. Inset is the SEM image focused on the etched edge.

In the experiment, two etching methods were applied: dry etching (via oxygen plasma) and wet etching (via hydrofluoric acid). Figure S3.2b compares the losses of the devices with the same etched depth ≈ 2.5 μm, for different process conditions. It shows that we can get etched chips of acceptable loss (less than 4 dB), via either dry etching or wet etching, but the coupler protection is necessary. Figure S3.2c shows the masks for the graphene-Al_2O_3-graphene structure fabrication. Patterns marked by I, II, III, IV are for lithography operations of the bottom layer graphene, bottom layer Au electrodes, top layer graphene, and top gate respectively. Figure S3.2d illustrates the resulting graphene Raman spectrum, before and after transfer onto the chip. Pumped with a 514 nm laser and after transfer, the graphene defect *D* peak is

negligible, the *G* peak width is ≈ 6 cm^{-1}, and the *2D* peak width is ≈ 14. Intensity ratio of *G* to *2D* is ≈ 0.75. The Raman spectra are comparable to that of monolayer and dual-layer CVD graphene measured during our fabrication.

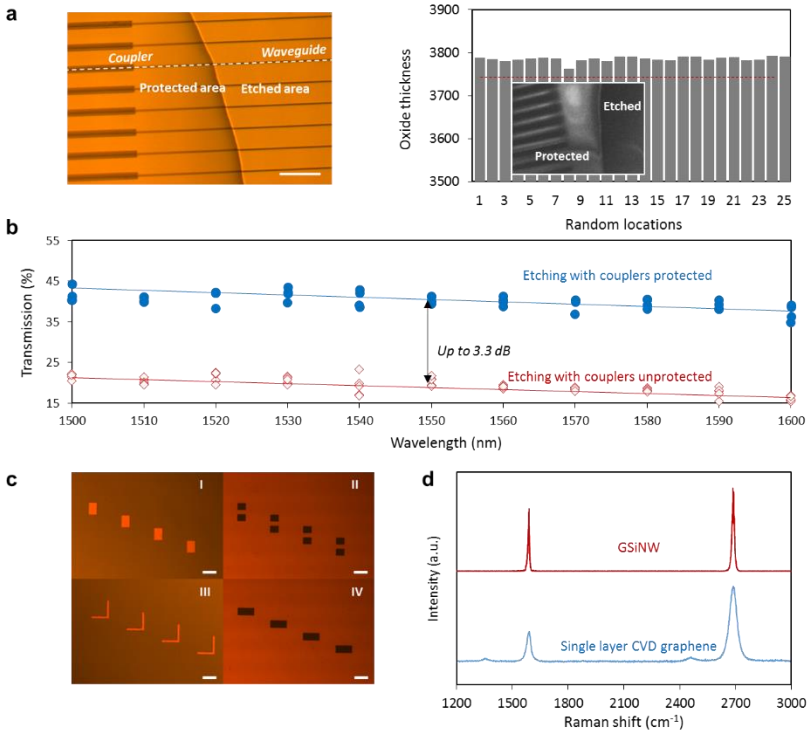

Figure S3.2 | Chip processing for dual-layer graphene interaction. a, (Left) Optical micrograph of the processed chip, with arrayed input-output straight waveguides and inverse couplers. Brighter area is etch-controlled down to 100 nm for the graphene-nitride interaction. Scale bar: 100 μm. (Right) Etched thickness. **b,** Average oxide thickness after etching, remaining <50 nm oxide upon the core . **c,** Designed masks for the source-drain-gate implementation. Scale bar: 150 μm. **d,** Normalized graphene Raman spectra, before and after GSiNW transfer and electrodes processing. Blue curve: single layer CVD graphene; red curve, the GSiNW.

S4. Experimental architecture

S4.1 Experimental setup

Figure S3.1 illustrates the experimental setup. A mode-locked pump pulse is launched into the GSiNW from the left, while an amplified continuous-wave (CW) signal is counter-launched from the right, both in TM polarizations. In the DFG process, the energy of the converted signal photon arises from the pump photon less the plasmon energy - the generation of the plasmons could thus be observed by monitoring the transmitted signal intensity on the left output. To directly detect the DFG plasmon signal over the noise, we implement a 100 kHz modulation of the mode-locked 39.1 MHz pump laser with lock-in filtering and amplification, along with balanced detection. To enable the detection of DFG plasmon signal, the launched light beams are TM polarized. A high power mode-locked laser is applied the pump and a

pre-amplified CW tunable laser is applied as the signal. Balanced photodetection (BPD) and lock-in amplification are implemented to extract the small plasmon signal from white noise exactly and clearly.

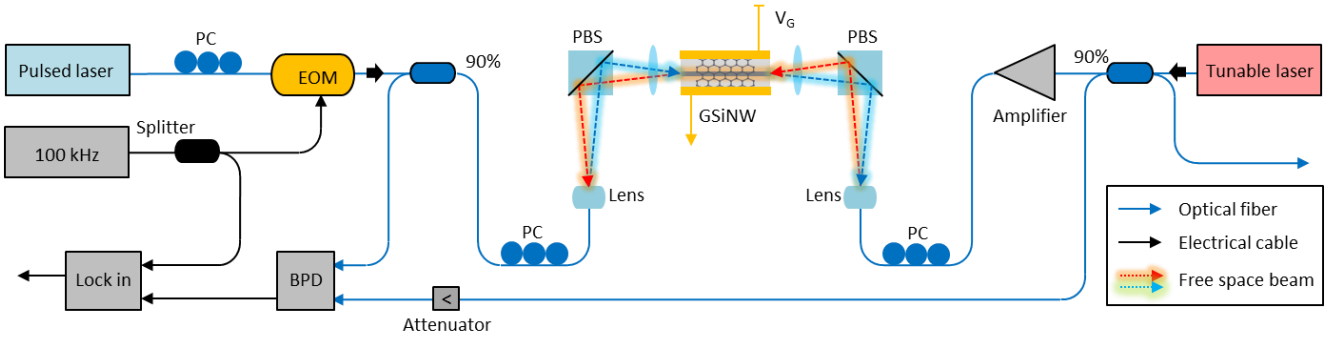

Figure S4.1 | Experimental setup. Measurement setup: A mode-locked picosecond laser serves as the ~ 400 pJ pump (f_p = 195.8 THz, 2.2 ps pulse duration, 39.1 MHz repetition rate, and 200 W peak power), which is slowly modulated at 100 kHz sinusoidally for single phase lock-in detection. A broadband tunable CW laser serves as the signal frequency f_s, amplified in the 1570 nm to 1610 nm band up to 1.6 W. The plasmon signal is detected in a balanced photodetector, with lock-in detection.

S4.2 Pulsed pump and its modulation

Here we use a nonlinear process to detect the in-plane graphene plasmons. To enhance the DFG nonlinear signal detection, a mode-locked picosecond fiber laser serves as the pump, which is pre-filtered and modulated. The spectral and temporal profile of the near-transform-limited pump launched onto the chip is illustrated in Figure S4.2. The spectrum is centered at 1531.8 nm (195.8 THz) with an ≈ 0.7 nm linewidth (Figure S4.2a). Figure S4.2b shows the temporal profile with 2.2 ps full-width half-maximum, measured by frequency-resolved optical gating (FROG), with a maximum average power of 16.1 dBm (40.7 mW) at 1531.9 nm and a quasi-linear increase (Figure S4.2c). Figure S4.2d shows the modulated pulsed pump, with the slow 100 kHz envelope and the embedded 39.1 MHz pulses inside. Figure S4.2e shows the corresponding electronic spectrum.

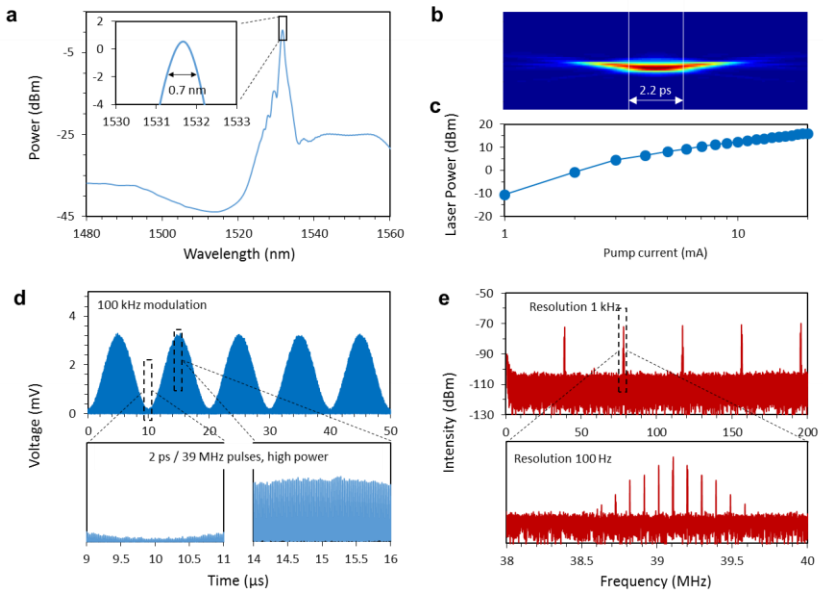

Figure S4.2 | Pump filtering and modulating. a, Spectra of the ps pulsed pump, with 0.7 nm spectral linewidth. **b,** Pulsewidth of the pulsed pump, measured by using FROG. **c,** Measured averaged power of the ps pulsed pump. **d,** Temporal profile of the pump, modulated by a 100 kHz sinusoidal signal for lock-in amplification, with the embedded 39.1 MHz pulses inside. **e,** Corresponding electronic spectrum of the modulated pump, 39.1 MHz peaks with 100 kHz harmonics are clear.

S4.3 CW signal light balanced detection and locked-in amplification

Figure S4.3a illustrates the setup to detect the weak DFG (with a 100 kHz modulation) from a strong background signal (CW) schematically. A CW tunable laser with intensity A is divided to be two paths. One passes the GSiNW while the other one serves as a reference. Then the DFG enhanced path with a 100 kHz gain is balanced by the reference, eliminating the CW component A. The dynamic intensity of the balanced signal has both gain and noise components. To extract the gain from noise, we use a lock-in amplifier at 100 kHz clock. Here, $N_{1,2,3}$ denote the noises and a, b, c are the attenuation and amplification factors. Correspondingly Figure S4.3b compares the measured intensities of CW signal (the DC component), the noise, the newly generated signal (from the DFG process, with 100 kHz oscillation), and the SNR before the balanced photodetector (BPD) (P1), after the BPD (P2), and after the lock-in amplifier (P3). We demonstrate that the BPD is predominantly used for DC balancing while the lock-in amplifier is applied to lock and amplify the 100 kHz gain. Figures S4.3c and S4.3d show the gating of the chip-scale GSiNWs with the micro-probes. V_G is tuned up to ± 4 V with 10 mV accuracy. Figure S4.3e illustrates the measured hysteresis loop of the GSiNW with V_{SD} at 10 mV. When $V_G \approx 0.25$ V, the bottom layer graphene approaches the Dirac point.

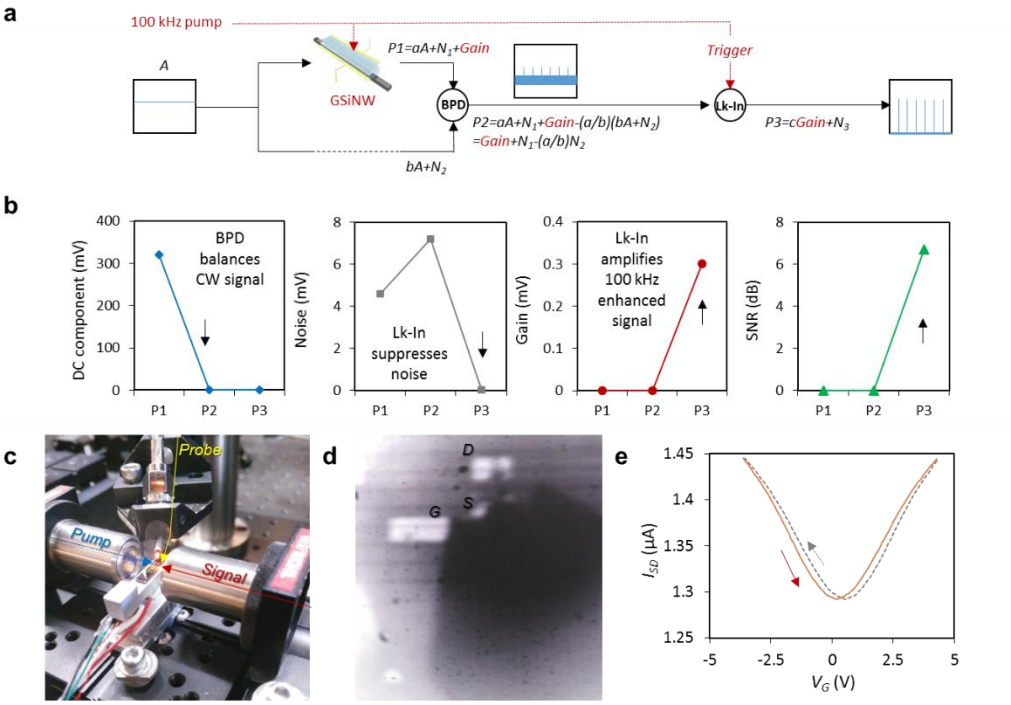

Figure S4.3 | Signal balanced detection and locked-in amplification. a, Schematic of the CW signal processing. Lk-In: lock-in amplification. **b,** Comparisons: DC component (original CW light), noise component, 100 kHz gain, and the resulting SNR. Here *P1,2,3* corresponds to the tap-out points in **a**. **c** and **d,** Gating the graphene-based semiconductor chip. **e,** Hysteresis loop of the GSiNW with V_G from -4 to 4 V (red) and 4 to -4 V (grey).

S5. Additional and supporting measurements

S5.1 Transmission of the GSiNW

Figure S5.1a shows the chip-scale 1500 to 1600 nm normalized transmitted spectrum, before and after covering with the graphene-Al_2O_3-graphene hybrid layer. The transmission of the nitride waveguide before etching is normalized as 0 dBm, and the launched power is ≈ 1 mW (significantly lower than the graphene saturated threshold). The initial 3.4 dB insertion loss is from the wet-etch chip processing; graphene coverage subsequently brings additional loss due to its monolayer broadband optical absorption. The loss of the shorter wavelengths is lower (red curve), perhaps due to the better mode field confinement. The graphene induced loss is ≈ 7.3 dB at 1500 nm (0.09 dB/μm), ≈ 8.3 dB at 1550 nm (0.1 dB/μm), and ≈ 9.5 dB at 1600 nm (0.12 dB/μm). Figure S5.1b tables the pump-signal polarization combinations – only when both the pump and signal are of TM polarization can DFG and the resulting THz plasmons be excited.

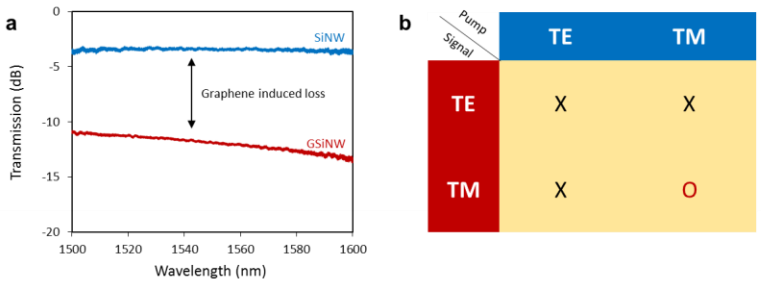

Figure S5.1 | Transmission and polarization combinations. a, Continuous-wave signal transmission: silicon nitride waveguide (SiNW) without graphene layers (blue curve), and the GSiNW (red curve). **b,** When pump and signal varies their polarizations, only TM-TM can generate the DFG-based plasmon in graphene.

S5.2 Pre-saturation of the GSiNW by using CW signal

Figure S5.2a plots the transmission versus the launched power, over four GSiNW samples. Red dots are the measurements with theoretically fitted blue curve and the noise width is denoted by the grey region. Clear saturable absorption of the GSiNW starts from ≈ 100 mW (20 MW/cm^2) and the GSiNW is almost fully saturated when the launched power is above 1 W (0.2 GW/cm^2). The saturable absorption induced transmission increase is ≈ 63%. Enabled by the saturable absorption [S46], the high power pulsed pump can modulate the low power CW signal.

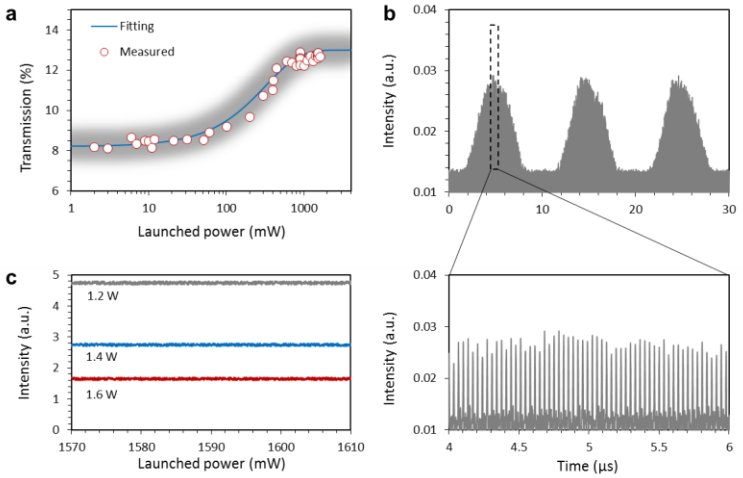

Figure S5.2 | Saturable absorption induced modulation. a, Saturable absorption of the GSiNW. **b,** Modulated signal. **c,** Modulation enhanced 100 kHz signal, amplified by the lock-in amplifier, when the CW signal is 1 W (grey), 1.2 W (blue) and 1.4 W (red).

Figure S5.2b shows the modulated CW signal measured after the balanced photodetector. The launched CW signal and pump powers are 1 mW and 32 mW respectively. The modulated CW is of the same temporal profile and the same repetition rate of the pulsed pump. Hence, the lock-in amplifier cannot filter off the modulation induced signal enhancement. That means, after the lock-in amplifier, the

background of the enhanced signal spectrum is not 0. For pristine graphene, the modulation can be three orders of magnitude larger than the DFG based enhancement. When the modulation is too large, it might saturate the detector, rendering the DFG enhanced peak undetectable. To suppress this modulation, we use high power CW signal to *pre-saturate* the graphene layers. Figure S5.2c shows the lock-in amplified signals, by using the CW laser with 1.2 W, 1.4 W, and 1.6 W powers.

S5.3 DFG enhanced signal of a GSiNW with 60 nm thick Al_2O_3

The Al_2O_3 layer thickness not only determines the V_G-I_{SD} curve of the graphene-Al_2O_3-graphene transistor, but also influences the plasmon coupling. Figure S5.3 shows the DFG enhanced signal at V_G = 0 V, when thickness of the Al_2O_3 is 60 nm. Compared to the GSiNW with 30 nm thick Al_2O_3 (blue curve, peak location 1593.7 nm, f_{SP} = 7.4 THz), the GSiNW with 60 nm thick Al_2O_3 (red curve) has a peak location at 1589.9 nm (f_{SP} = 7.1 THz). We regard that there is little plasmon coupling in a 60 nm graphene-Al_2O_3-graphene system.

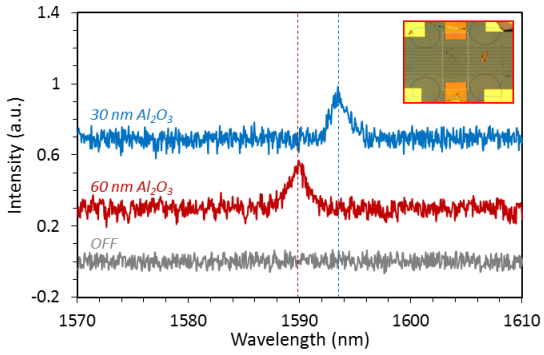

Figure S5.3 | Enhanced spectra. a, Grey: pump off; Blue: GSiNW with 30 nm Al_2O_3; Red: GSiNW with 60 nm Al_2O_3. Inset: Optical micrograph of the GSiNW with 60 nm Al_2O_3.

S5.4 Measurement of the plasmons with pump frequency tuning

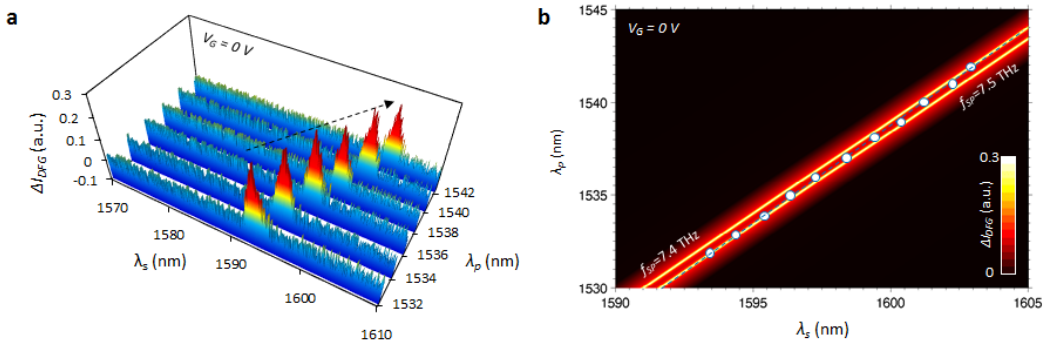

Figure S5.4 | Tuning the pump frequency. a, Spectra of the enhanced signal, when λ_p is tuned from 1532 nm to 1542 nm. **b,** Measured λ_p-λ_s correlation, with the f_{SP} slightly shifted from 7.5 to 7.4 THz due from dispersion matching. When λ_p is 1532 nm, f_{SP} is 7.5 THz; when λ_p is 1542 nm, f_{SP} is 7.4 THz.

When λ_p is tuned from 1532 nm to 1542 nm (195.8 THz to 194.6 THz), the enhanced signal peak λ_s is shifted from 1593.2 nm to 1603 nm (188.3 THz to 187.2 THz) as shown in Figures S5.4a and S5.4b. During this process, f_{SP} decreases from 7.5 THz to 7.4 THz. The trace of the f_{SP} follows the graphene plasmonic dispersion well, as described in the main text.

Supplementary References:

S1. Low, T. & Avouris, P. Graphene plasmonics for terahertz to mid-infrared applications. *Nano Lett.* **8,** 1086–1101 (2014).

S2. Ferguson, B. & Zhang, X. Materials for terahertz science and technology. *Nat. Mat.* **1,** 26-33 (2002).

S3. Tonouchi, M. Cutting-edge terahertz technology. Nat. Photon. **1,** 97-105 (2007).

S4. Williams, B. Terahertz quantum-cascade lasers. *Nat. Photon.* **1,** 517-525 (2007).

S5. Li, J. et al. Electro-optical frequency division and stable microwave synthesis. *Science* **345,** 309-313 (2014).

S6. Bartalini, S. et al. Frequency-comb-assisted terahertz quantum cascade laser spectroscopy. *Phys. Rev. X* **4,** 021006 (2014).

S7. Belkin, M, et al. Terahertz quantum-cascade-laser source based on intracavity difference-frequency generation. *Nat. Photon.* **1,** 288-292 (2007).

S8. Lu, Q. et al. Room temperature continuous wave, monolithic tunable THz sources based on highly efficient mid-infrared quantum cascade lasers. *Sci. Rep.* **6,** 23595 (2016).

S9. Jung, S. et al. Broadly tunable monolithic room-temperature terahertz quantum cascade laser sources. *Nat. Comm.* **5,** 4267 (2014).

S10. Vijayraghavan, K. et al. Broadly tunable terahertz generation in mid-infrared quantum cascade lasers. *Nat. Comm.* **4,** 2021 (2013).

S11. Tian, Y. et al. Femtosecond-laser-driven wire-guided helical undulator for intense terahertz radiation. *Nat. Photon.* **11,** 242-246 (2017).

S12. Chen, J. et al. Optical nano-imaging of gate-tunable graphene plasmons. *Nature* **487,** 77-81 (2012).

S13. Fei, Z. et al. Gate-tuning of graphene plasmons revealed by infrared nano-imaging. *Nature* **487,** 82-85 (2012).

S14. Ni, G. Ultrafast optical switching of infrared plasmon polaritons in high-mobility graphene. *Nat. Photon.* **10,** 244–247 (2016).

S15. Ju, L. et al. Graphene plasmonics for tunable terahertz metamaterials. *Nat. Nanotech.* **6,** 630-634 (2011).

S16. Alonso-González, P. et al. Controlling graphene plasmons with resonant metal antennas and spatial conductivity patterns. *Science* **334,** 1369-1373 (2014).

S17. Cox, J. & Abajo, F. Electrically tunable nonlinear plasmonics in graphene nanoislands, *Nat. Comm.* **5,** 5725 (2014).

S18. Nikitin, A. et al. Real-space mapping of tailored sheet and edge plasmons in graphene nanoresonators, *Nat. Photon.* **10,** 239-243 (2016).

S19. Yan, H. et al. Damping pathways of mid-infrared plasmons in graphene nanostructures. *Nat.*

Photon. **7,** 394-399 (2013).
S20. Constant, T. et al. All-optical generation of surface plasmons in graphene. *Nat. Phys.* **12,** 124-127 (2016).
S21. Phare, C., et al. Graphene electro-optic modulator with 30 GHz bandwidth. *Nat. Photon.* **9,** 511-514 (2015).
S22. Snyder, A. & Love, J. Optical waveguide theory. (Springer, USA, 1984).
S23. Vakil, A. & Engheta, N. Transformation optics using graphene. *Science* **332,** 1291-1294 (2011).
S24. Mikhailov, S. A. & Ziegler, K. New electromagnetic mode in graphene. *Phys. Rev. Lett.* **99,** 016803 (2007).
S25. Hanson, G. Dyadic Green's functions and guided surface waves for a surface conductivity model of graphene. *J. App. Phys.* **103,** 064302 (2008).
S26. Koppens, F., Chang, D. & Abajo, F. Graphene Plasmonics: A platform for strong light matter interactions. *Nano Lett.* **11,** 3370-3377 (2011).
S27. Holmes, W. et al. Measurements of thermal transport in low stress silicon nitride films. *Appl. Phys. Lett.* **72,** 2250-2252 (1998).
S28. Hwang, E., Sensarma, R. & Sarma, S. Plasmon-phonon coupling in graphene. *Phys. Rev. B.* **82,** 195406 (2010).
S29. Zhu, W. et al. Silicon nitride gate dielectrics and band gap engineering in graphene layers. *Nano Lett.* **10,** 3572-3576 (2010).
S30. Luxmoore, I. et al. Strong coupling in the far-infrared between graphene plasmons and the surface optical phonons of silicon dioxide. *ACS Photon.* **1,** 1151−1155 (2014).
S31. Brar, V. et al. Electronic modulation of infrared radiation in graphene plasmonic resonators. *Nat. Comm.* **6,** 7032 (2015).
S32. Wunsch, B. et al. Dynamical polarization of graphene at finite doping. *New J. Phys.* **8,** 318-326 (2006).
S33. Jablan, M., Buljan, H. & Soljacic. M. Plasmonics in graphene at infrared frequencies. *Phys. Rev. B.* **80**, 245435 (2009).
S34. Mikhailov, S. A. Theory of the giant plasmon-enhanced second-harmonic generation in graphene and semiconductor two-dimensional electron systems. *Phys. Rev. B.* **84,** 045432 (2011).
S35. Yao, X., Tokman, M. & Belyanin, A. efficient nonlinear generation of THz plasmons in graphene and topological insulators. *Phys. Rev. Lett.* **112,** 055501 (2014).
S36. Hwang, E. & Sarma, S. Plasmon modes of spatially separated double-layer graphene. *Phys. Rev. B* **80,** 205405 (2009).
S37. Profumo, R. et al. Double-layer graphene and topological insulator thin-film plasmons. *Phys. Rev. B* **85,** 085443 (2012).
S38. Wang, B. et al. Optical coupling of surface plasmons between graphene sheets. *Appl. Phys. Lett.* **100,** 131111 (2012).
S39. Stauber, T. Plasmonics in Dirac systems: from graphene to topological insulators. *J. Phys. Condens. Matter.* **26**, 123201 (2014).
S40. Das, A. et al. Monitoring dopants by Raman scattering in an electrochemically top-gated graphene transistor. *Nat. Nanotech.* **3,** 210-215 (2008).

S41. Gu, T. et al. Regenerative oscillation and four-wave mixing in graphene optoelectronics. Nat. Photon. **6,** 554-559 (2012).
S42. Hendry, E. et al. Coherent nonlinear optical response of graphene. *Phys. Rev. Lett.* **105**, 097401 (2010).
S43. Palomba, S. & Novotny, S. Nonlinear excitation of surface plasmon polaritons by four-wave mixing. *Phys. Rev. Lett.* **101,** 056802 (2008).
S44. Simkhovich, B. & Bartal, G. Plasmon-enhanced four-wave mixing for superresolution applications. *Phys. Rev. Lett.* **112,** 056802 (2014).
S45. Sun, Y., Qiao, G. & Sun, G. Direct generation of graphene plasmonic polaritons at THz frequencies via four wave mixing in the hybrid graphene sheets waveguides. *Opt. Exp.* **22,** 27880-27891 (2014).
S46. Sun, Z., Martinez, A. & Wang, F. Optical modulators with 2D layered materials. *Nat. Photon.* **10,** 227–238 (2016).