**Response to the referee's comments**:

We are very grateful to the referee for the helpful comments on our manuscript. The referee suggests a few minor revisions. (The referee says: "I think the paper is well written and the main result is useful. The careful MCMC study adds reliability to the constraints. Therefore, I am happy to recommend publication after the authors make a few small revisions.") According to the referee's comments, we have made revision, and we believe that the paper has been significantly improved. In the following we shall respond to the referee's comments and describe the changes in detail.

(1) Comment: It may be useful to describe in a bit more detail how the theoretical predictions were made. For example, the sterile neutrinos used in this study have masses of order 0.1-1eV, and does this mean that their perturbations were included in the numerical analysis? Also, it is not very clear how the parameter gamma is used in the joint MCMC analysis. In principle, the whole analysis can be carried out without including gamma and using only the 6 base parameters plus neutrino parameters. So how does gamma affect the theoretically calculated values of the different cosmological observables such as WL and CMB?

Response: We thank the referee for this good question. We think that the essence of this question is to ask how the additional parameter $\gamma$ is included as a free parameter in the numerical calculations. We know that in GR the linear growth factor $D(t)$ obeys the differential equation (on scales much smaller than the horizon),

$$\ddot{D}_{\rm GR} + 2H(z)\dot{D}_{\rm GR} - \frac{3}{2}\Omega_m H_0^2(1+z)^3 D_{\rm GR} = 0.$$

The solution to this equation can be written in integral form for specific forms of $H(z)$. To a very good approximation, the growth rate of linear perturbations in GR is

$$f_{\rm GR} \equiv \frac{d\ln D_{\rm GR}}{d\ln a} \approx [\Omega_m(z)]^{0.55}.$$

In this work, to test GR, we use the parametrization $f(a) = \Omega_{\rm m}(a)^\gamma$ to describe the density perturbations in the $\Lambda$CDM cosmology, and thus we introduce an extra parameter $\gamma$ into the model. With the linear growth factor $D(a)$ in hand, we can calculate the density perturbations in late-time universe for all the components.

To respond to the referee's comment, we add a paragraph to describe how to include the additional parameter $\gamma$ in the calculations (Page 2, Sec. II): "To constrain the growth index $\gamma$, we use the parametrization $f(a) = \Omega_{\rm m}(a)^\gamma$ to describe the density perturbations in the $\Lambda$CDM cosmology, and thus we introduce an extra parameter $\gamma$ into the model. We use the RSD measurements of $f(z_{\rm eff})\sigma_8(z_{\rm eff})$ to set constraints on $\gamma$. We follow the procedure of Sec. 9.1 in Ref. [23] to include $\gamma$ as an additional parameter. Here, it is helpful to briefly describe how the parameter product $f_\gamma(z_{\rm eff})\sigma_{8,\gamma}(z_{\rm eff})$ is derived in the theoretical calculations by the following two steps: (i) Since in this description the value of $\sigma_8(z_{\rm eff})$ depends on $\gamma$, we have to recalculate this value by using the parametrization of $f_\gamma(a_{\rm eff}) = \Omega_{\rm m}(a_{\rm eff})^\gamma$. We first calculate the growth factor, $D(a_{\rm eff}) = \exp\left[-\int_{a_{\rm eff}}^{1} da' f(a')/a'\right]$, where $a_{\rm eff}$ is the scale factor at the effective redshift $z_{\rm eff}$. Then, we derive $\sigma_{8,\gamma}(z_{\rm eff})$ by the extrapolation from the matter dominated epoch to the effective redshift, $\sigma_{8,\gamma}(z_{\rm eff}) = \frac{D_\gamma(z_{\rm eff})}{D(z_{\rm hi})}\sigma_8(z_{\rm hi})$, where $\sigma_8(z_{\rm hi})$ is calculated at $z_{\rm hi} = 50$ which is in the deep matter-dominated regime, where $f(z) \approx 1$. (ii) We calculate the growth rate by using the parametrization $f_\gamma(z_{\rm eff}) = \Omega_{\rm m}(z_{\rm eff})^\gamma$. Thus, now, we can obtain the parameter product $f_\gamma(z_{\rm eff})\sigma_{8,\gamma}(z_{\rm eff})$ in the numerical calculations."

(2) Comment: I think it is an overly strong claim to say that this study finds no hint of the existence of light sterile neutrinos. What it really shows is that adding such particles does not completely erase the tension.

Response: We agree with the referee. We have removed this strong claim in the revised version.

(3) Comment: Tensions of various sigmas are mentioned in this paper, so it is helpful to define it in a prominent place somewhere.

Response: To answer this question, we have added a paragraph in the revised version (Page 3, the last paragraph of Sec. II): "In this paper, tensions between different observations for some cosmological parameters will occasionally be mentioned, so it is helpful to clearly describe how to estimate the degree of tension between two observations for some parameter in this place. Assume that, for a parameter $\xi$, we have its 68% confidence level ranges $\xi \in [\xi_1 - \sigma_{1,\rm low},\ \xi_1 + \sigma_{1,\rm up}]$ from an observation (O1) and $\xi \in [\xi_2 - \sigma_{2,\rm low},\ \xi_2 + \sigma_{2,\rm up}]$ from another observation (O2). The statement that "the tension between O1 and O2 is at the $a\sigma$ level" means that we have $a = (\xi_2 - \xi_1)/\sqrt{\sigma_{2,\rm low}^2 + \sigma_{1,\rm up}^2}$ for the case $\xi_2 > \xi_1$, and vice versa. In this work, we estimate the degree of tension between different observations by this simple way."

In addition, we have corrected some minor errors and typos, and we have also added several references (see Refs. [40,50,74-77,82-90]).

It can be seen that we have tried our best to revise our manuscript following the referee's comments. We wish that our revision could make the referee satisfactory and warrant the publication of this paper in Physics Letters B. Thanks a lot!