Peer Review File

Manuscript Title: Detection of stellar light from quasar host galaxies at redshifts above 6

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Dear Editor,

I have read with interest the paper by Ding et al. on the "First deetctions of stellar light from quasar host galaxies at z>6"

The paper presents JWST imaging data of two quasars at z>6 obtained with NIRCam in two filters. After careful subtraction of the PSF, the authors obtain clear evidence for extended emission in the longer band (F356W) for both objects, while they obtain a fainter and more marginal detection in the bluer band (F150W) only for one of the two targets.

The results are interesting.
However, I think that the interpretation and conclusions are not supported by the data.

The first important issue is that the filter F365W is most heavily contaminated, and very likely dominated, by [OIII]+Hbeta emission at this redshift. Galaxies at these early times can have very high EW of these lines, and especially if ionization by AGN is involved.
An example of this effect can be easily seen in the several strong [OIII]+Hb detections by the recent EIGER results (Kashino et al. 2022, Matthee et al. 2022).
They have shown that such strong lines can easily dominate the emission in this filter at high redshift.
If so, the data presented in the paper may reval nebular emission in the quasar host galaxy, which can possibly be excited by star formation, but can even result from the quasar excitation, i.e. it could be simply tracing the quasar's Narrow Line Region.

Similarly, the more marginal detection of extended emission in the F150W band in the second target could be associated with extended CIII] emission of the NLR or from star formation in the host galaxy. Extended CIII] emission has been observed in lower redshift quasars (e.g. Guo et al. 2020).

Unless the authors can support their claims that the detected extended emission is dominated by stellar light, e.g. by supplementing the data with IFU spectroscopic observations, I think that their interpretation is not tenable.

Even if they could somehow prove that the observed extended emission is dominated by stellar light, inferring stellar masses with only two filters is essentially impossible. With only two photometric points the stellar population properties are totally degenerate with respect to any possible combination of mass, SFR, age, metallicity, and dust attenuation. The errors on the stellar masses provided by the authors are totally unrealistic. As a minimum, the authors should at least obtain photometry in two additional filters.

Given the two major issues outlined above, I'm afraid I cannot recommend the paper for publication.

Referee #2 (Remarks to the Author):

The detection of host star light in z>6 quasars has been long awaited. Undoubtedly JWST is required to make such challenging measurements, and the current paper represents the beginning of an exciting era. The design, goals and timely analyses of this project are all excellent. As such, I am very much in favor of the publication of these results in Nature. However, I believe the current manuscript should and can be significantly improved both in the scientific scope and in the technical verification. In my opinion, the paper is a little rushed, and does not capitalize on the available data. Below I will first describe some major concerns, followed by a list of minor suggestions. I hope that with these improvements (additional analyses required) the conclusions from this paper will be strengthened, and the scope broadened to set this paper as a milestone in z>6 quasar host studies.

* Scope

The primary interest of detecting z>6 quasar host star light is to study the relation between stellar mass and black hole mass, which may shed light on the question why these black holes can grow so massive and so fast in the early Universe. This is a major problem in this field, as highlighted in the summary paragraph of this paper. As such, I expect the first paper on host detection with JWST would at least discuss the host stellar masses in the context of the Mbh-M* relation. The NIRSpec data for measuring BH masses are already taken and should be included in this paper. From the discussion in Method, I understand that there are already preliminary results on the NIRSpec data, but the authors decided to defer to a separate paper.

Including BH mass measurements and discussions on the Mbh-M* relation compared with low-z results would significantly increase the impact of this paper, and report a more complete story of high-redshift quasar hosts with JWST, as anticipated for Nature-grade papers. Without this critical piece of information, the current paper is still valuable, but the scope is much limited. Since this paper was already posted to arXiv, there is no concern of not getting the proper credit for the team. I therefore strongly advise the authors to include BH mass measurements and discuss them properly in the paper. There are many experts in the author list who can perform such BH mass measurements with NIRSpec data, and in my experience the spectral analysis is straightforward.

* Image decomposition and host measurements

The main results from this paper are the detection of host star light and constraints on the host stellar mass of two z>6 quasars. Additionally the paper reports morphological measurements of the two quasar hosts, finding them to be compact, disk-like, with potential displacement from the quasar light. The authors performed a series of tests to justify the methodology of their image decomposition and host measurements. My main concern here is that the issue of PSF mismatch in the image decomposition of the quasar point source + extended host is not adequately addressed, which I detail below.

First of all, there is still significant systematics regarding JWST NIRCam PSFs. For example, this paper (https://ui.adsabs.harvard.edu/abs/2022MNRAS.517..484N/abstract) suggests there are significant spatial variations across the FoV that can be up to ~15-20%. These spatial PSF variations can easily swamp the host signal or lead to artificial detection of extended emission around a point source. Indeed, the authors find not all stars in the FoV have the same PSF, and the reported best-fit model depends on the choice of the fiducial “PSF” model (either the PSF that achieves the smallest chi2, or some sort of averaged PSF). I am not convinced that using chi2 is necessarily the correct choice here, because a mismatched PSF could be compensated by artificial host emission, leading to better chi2 performance.

This problem is reflected in the sanity check shown in the bottom row of Fig. E3, where a randomly picked PSF star is used to fit another star. Because of the PSF mismatch, there is central residual after the fit. If the authors swap the two stars in the fitting, then they will likely find the residuals are in the broad wings, which I have verified with some simple simulations. In practice, the PSF mismatch is more complicated. For example, if two PSFs only differ slightly in the wing but are identical in the core, then using one of them (the one with a narrower wing) to fit the other would mainly lead to extended residuals. The PSF mismatch might also be non-axisymmetric.

Given these considerations, I recommend the authors to perform a more thorough PSF analysis to quantify the variations of PSF across the FoV and the resulting systematic uncertainties of the claimed host detection. I am sure that the authors can come up with a more sophisticated scheme, but below is a suggestion as a starting point.

First, I would like to see all the results of fitting stars with other stars in the same FoV and filter (each pair of stars are fit to each other). This would inform us how much PSF variation there is. The authors should quantify the range of variations with some metrics, e.g., the fractional flux of “extended” residuals to the total flux — but certainly there are other metrics that matter as well. By design half of these fitting pairs should result in extended residuals. Then based on these tests, the authors can assess the quality of their quasar+host decomposition and the significance of the host detection. If the range of results (as quantified by the defined metrics) for the quasar fits with all PSF stars is similar to that for the tests of stars fitting stars, then there is not much evidence for host detection of the quasar. However, if the authors find significantly different results when fitting the quasar with all PSF stars, then they may have a case of real host detection.

However, detecting the host emission does not mean that the host measurements must be accurate. This is a critical issue for reliable stellar mass estimates, as well as constraints on morphological properties of the host. My concern is that the systematic uncertainties of the host measurements are substantial, and currently under-appreciated in quasar host studies with JWST imaging. To fully address this issue, the best approach is to simulate NIRCam images of z=6 quasar+host systems with known input host parameters, and apply the PSF templates to quantify the accuracy of recovered host parameters (e.g., centroid, flux, size and Sersic index). This is a worthwhile exercise for this important paper on first detection of z~6 quasar hosts in star light.

* Minor suggestions

I will provide a more complete list of minor comments when this manuscript is in a more advanced stage. Below are some points that I currently have.

- In several places (including the summary paragraph) the authors claim that "there are no host detection in star light beyond z>2". I am not sure this is true. HST-WFC3 F160W is able to detect host starlight at z>2. There are papers on such measurements at z>2 (whether or not these host measurements are robust is a different question; but certainly some of them look better than the results in this paper), e.g., https://ui.adsabs.harvard.edu/abs/2015ApJ...806..218G/abstract, https://ui.adsabs.harvard.edu/abs/2022arXiv220911249C/abstract. If you don't rely on direct imaging, then SED fitting arguably can also derive host star light for z>2 quasars (e.g., https://ui.adsabs.harvard.edu/abs/2016ApJ...824...70D/abstract). So my suggestion is to simply modify this claim to refer to "z>4 quasars" across this paper.

- Somewhat related to my major concern above, I think most people working on quasar-host imaging decomposition would agree that it is very difficult to get the host morphology right in these quasar-light-dominated systems. I think the authors should emphasize this caveat more, and caution on the interpretation of some best-fit parameters.

- I am particularly skeptical about the claimed centroid offset of the hosts, unless the authors can demonstrate the robustness of the centroid constraints with simulated images + observed PSFs. I am also confused about the claim ".. similar offsets can be found in ~20% of lower redshift quasars [Ref 13,24 by the same primary authors]" with the same concern on the robustness of host centroid measurements. I would like to see more discussions on this, and citations to independent works supporting or against these low-z results.

- Fig. 3 lower panel: no error bars on the data points?

- Table 1: rows on positional offset (") and (kpc). I do not understand why the fractional errors (%) would be different between the angular and physical offsets.

- Methods. About background subtraction, how were the close neighbors to J2255 subtracted? Or were the two close neighbors part of the decomposition fitting?

- About the final fitting results with uncertainties. In Methods it states " We take the results from the five top-performing PSF models in the updated library. We determine our final result parameters with uncertainties by weighting their chi2 values." How exactly is this done? E.g., is the best-fit parameter the 1/chi2-weighted mean of the five models? What about the uncertainties? Are the uncertainties marginalized?

- Methods on “Stellar mass and luminosity of the host galaxy”. NIRSpec data are mentioned here to justify the neglecting of line emission in the broad-band photometry. This is another (minor) reason why these NIRSpec data should be published along with this paper.

- A discussion on the systematic uncertainties on the stellar mass estimates from SED modeling would be useful.

Author Rebuttals to Initial Comments:

Your manuscript entitled "First detections of stellar light from quasar host galaxies at z > 6" has now been seen by 2 referees, whose comments are attached below. While they find your work of potential interest, as indeed do we, they have raised substantial concerns that first need to be addressed before we can consider the paper further for possible publication in Nature.

Both referees raise what appear to be substantial concerns. As an aside, I will note that, given the changing calibrations for the JWST pipelines, most papers are encountering difficulties with the referees.

We are very grateful for the constructive feedback provided by the editor and reviewers, which helped us to enhance the quality of this paper to the level expected by Nature.

Regarding the current JWST calibration issue, as now explained in the paper, the remaining calibration uncertainties contribute 0.1 mag (0.04 dex) or less to the host stellar mass uncertainty, which is a small contribution to our error budget. Our final goal is to compare these measurements with the inferred black hole masses, which were measured using the single-epoch methods with uncertainty at the 0.4 dex level.

We have carefully considered and addressed all the points raised by the referees and revised the paper accordingly. Most significantly, the paper now includes our JWST NIRSpec results, allowing us to quote accurate black hole masses. To make room for this, we have dropped the discussion of the size-luminosity relation of quasar hosts, leaving that for a future paper. All the changes to the text are highlighted in blue color.

Response to the Referee #1 (Remarks to the Author):

The results are interesting. However, I think that the interpretation and conclusions are not supported by the data.

Reply: We thank the referee for raising these reasonable concerns. We have addressed them in the paper. We explain the changes in detail below.

The first important issue is that the filter F365W is most heavily contaminated, and very likely dominated, by [OIII]+Hbeta emission at this redshift. Galaxies at these early times can have very high EW of these lines, and especially if ionization by AGN is involved. An example of this effect can be easily seen in the several strong [OIII]+Hb detections by the recent EIGER results (Kashino et al. 2022, Matthee et al. 2022). They have shown that such strong lines can easily dominate the emission in this filter at high redshift. If so, the data presented in the paper may reveal nebular emission in the quasar host galaxy, which can possibly be excited by star formation, but can even result from the quasar excitation, i.e. it could be simply tracing the quasar's Narrow Line Region.

Reply: We now include NIRSpec observations in this paper. With this data, we show that [OIII] + narrow Hb does not dominate the inferred light from the host. For J2236, the slit was fortuitously aligned along the major axis of the galaxy (slit position now shown in the F356W images of Fig. 2) thus most of the host starlight falls within the slit. We perform a multi-component (broad+narrow) emission line fitting to the NIRSpec spectrum and find that the entire [OIII] + narrow Hb emission contributes less than ~3% of the NIRCam's host detection. Highly supportive of the above and described on Page 6 of the current manuscript, the

NIRSpec spectrum of J2236 shows strong stellar absorption lines (Hγ, Hδ, Hε, and CaII H+K), which supports the host detection in the NIRCam images. The host contribution to the total spectral light is at the level found with NIRCam.

Regarding J2255, the spatial location of the slit is not as optimal as above. A simple assessment of the line flux shows that it is comparable to the NIRCam measurement of the host flux. However, the spatial distribution of Hb+[OIII] emission is very compact thus most of the flux is removed with the decomposition of the NIRCam data. We perform with the NIRSpec data a careful subtraction of the unresolved emission to reveal the extended Hb+[OIII] component, which results in a contribution of ~15% of the host galaxy flux as being attributed to line emission. (see Methods, last section for details).

Similarly, the more marginal detection of extended emission in the F150W band in the second target could be associated with extended CIII] emission of the NLR or from star formation in the host galaxy. Extended CIII] emission has been observed in lower redshift quasars (e.g. Guo et al. 2020).

Reply: Given the high stellar mass and low level of [OIII]+narrow HB for J2236, it's very unlikely that CIII] lines dominate the extended emission.

Unless the authors can support their claims that the detected extended emission is dominated by stellar light, e.g. by supplementing the data with IFU spectroscopic observations, I think that their interpretation is not tenable.

Reply: As described above, we have used the NIRSpec Fixed-Slit observations for this purpose. We have convinced ourselves that the stellar continuum detections are robust. We hope that it's now convincing to the referee.

Even if they could somehow prove that the observed extended emission is dominated by stellar light, inferring stellar masses with only two filters is essentially impossible. With only two photometric points the stellar population properties are totally degenerate with respect to any possible combination of mass, SFR, age, metallicity, and dust attenuation. The errors on the stellar masses provided by the authors are totally unrealistic. As a minimum, the authors should at least obtain photometry in two additional filters.

Reply: We agree that the limitation of two filters dominates the uncertainty of our stellar mass measurements. However, it is still possible to use two filters to constrain the stellar SED if these filters are carefully selected. For our case, we use F150W and F356W, on opposite sides of the galaxy 4000 □ break; this maximizes the sensitivity to the total stellar mass with a limited number of filters (as done in Bell & de Jong 2001). To demonstrate this, we have updated our SED fit by setting the parameters of the stellar populations (age, metallicity, dust attenuation, and strength of the emission lines) as free parameters and performed the MCMC analysis. We limit these parameters within the physical ranges and use Figure E4 to show the final result: the inferred stellar mass is not sensitive to age or metallicity but is sensitive to the Av value. With a sensible range for Av ([0.3, 1.0]), we have confirmed that our stellar masses can be measured to < 0.2 dex, probably aided by the more massive galaxies found here than reported at z > 6 in the literature. We recognize that there are systematics with the stellar mass estimates that we cannot address with this study (e.g., IMF). Also, this uncertainty level is smaller than the intrinsic scatter of our MBH estimate (0.4 dex) thus enabling a meaningful comparison between them.

Referee #2 (Remarks to the Author):

The detection of host star light in z>6 quasars has been long awaited. Undoubtedly JWST is required to make such challenging measurements, and the current paper represents the beginning of an exciting era. The design, goals and timely analyses of this project are all excellent. As such, I am very much in favor of the publication of these results in Nature. However, I believe the current manuscript should and can be significantly improved both in the scientific scope and in the technical verification. In my opinion, the paper is a little rushed, and does not capitalize on the available data. Below I will first describe some major concerns, followed by a list of minor suggestions. I hope that with these improvements (additional analyses required) the conclusions from this paper will be strengthened, and the scope broadened to set this paper as a milestone in z>6 quasar host studies.

Reply: We greatly thank the referee for the positive feedback and suggestions, which really have strengthened this paper. All of the comments and suggestions have been properly addressed, as described below. Let us add that the quick analysis and publication of the first version was the result of much preliminary preparation before the data arrived. The new simulations, inspired by your comments, could be completed on a month time scale since they were already available and used in the original proposal.

* Scope

The primary interest of detecting z>6 quasar host star light is to study the relation between stellar mass and black hole mass, which may shed light on the question why these black holes can grow so massive and so fast in the early Universe. This is a major problem in this field, as highlighted in the summary paragraph of this paper. As such, I expect the first paper on host detection with JWST would at least discuss the host stellar masses in the context of the Mbh-M* relation. The NIRSpec data for measuring BH masses are already taken and should be included in this paper. From the discussion in Method, I understand that there are already preliminary results on the NIRSpec data, but the authors decided to defer to a separate paper.

Including BH mass measurements and discussions on the Mbh-M* relation compared with low-z results would significantly increase the impact of this paper, and report a more complete story of high-redshift quasar hosts with JWST, as anticipated for Nature-grade papers. Without this critical piece of information, the current paper is still valuable, but the scope is much limited. Since this paper was already posted to arXiv, there is no concern of not getting the proper credit for the team. I therefore strongly advise the authors to include BH mass measurements and discuss them properly in the paper. There are many experts in the author list who can perform such BH mass measurements with NIRSpec data, and in my experience the spectral analysis is straightforward.

Reply: We have followed the referee’s suggestion to include the NIRSpec results in this paper, allowing us to measure Hbeta-based Mbh and study the Mbh-M* relation for the first time at z>6. Note that we now cross-reference the imaging and spectroscopic data for the two quasars to measure the properties of the host galaxies and the central BHs. See the Methods sections

“BH mass estimates” and “Emission from Hb+[OIII] in the hosts” for details. To make room for this, we have dropped the description of the AGN host size-luminosity relation.

* Image decomposition and host measurements

The main results from this paper are the detection of host star light and constraints on the host stellar mass of two z>6 quasars. Additionally the paper reports morphological measurements of the two quasar hosts, finding them to be compact, disk-like, with potential displacement from the quasar light. The authors performed a series of tests to justify the methodology of their image decomposition and host measurements. My main concern here is that the issue of PSF mismatch in the image decomposition of the quasar point source + extended host is not adequately addressed, which I detail below.

First of all, there is still significant systematics regarding JWST NIRCam PSFs. For example, this paper:
(https://ui.adsabs.harvard.edu/abs/2022MNRAS.517..484N/abstract) suggests there are significant spatial variations across the FoV that can be up to ~15-20%. These spatial PSF variations can easily swamp the host signal or lead to artificial detection of extended emission around a point source. Indeed, the authors find not all stars in the FoV have the same PSF, and the reported best-fit model depends on the choice of the fiducial “PSF” model (either the PSF that achieves the smallest chi2, or some sort of averaged PSF). I am not convinced that using chi2 is necessarily the correct choice here, because a mismatched PSF could be compensated by artificial host emission, leading to better chi2 performance.
Reply: We agree that more extensive tests of the PSF model were needed, including the effects of mismatched PSFs which are the dominant systematic in the measurement of host properties. We have followed the referee’s suggestions and added a number of additional sanity checks to validate the accuracy of our host measurements. They are described in the Methods section and are outlined below.

This problem is reflected in the sanity check shown in the bottom row of Fig. E3, where a randomly picked PSF star is used to fit another star. Because of the PSF mismatch, there is central residual after the fit. If the authors swap the two stars in the fitting, then they will likely find the residuals are in the broad wings, which I have verified with some simple simulations. In practice, the PSF mismatch is more complicated. For example, if two PSFs only differ slightly in the wing but are identical in the core, then using one of them (the one with a narrower wing) to fit the other would mainly lead to extended residuals. The PSF mismatch might also be non-axisymmetric.

Given these considerations, I recommend the authors to perform a more thorough PSF analysis to quantify the variations of PSF across the FoV and the resulting systematic uncertainties of the claimed host detection. I am sure that the authors can come up with a more sophisticated scheme, but below is a suggestion as a starting point.

First, I would like to see all the results of fitting stars with other stars in the same FoV and filter (each pair of stars are fit to each other). This would inform us how much PSF variation there is. The authors should quantify the range of variations with some metrics, e.g., the fractional flux of “extended” residuals to the total flux — but certainly there are other metrics that matter as well. By design half of these fitting pairs should result in extended residuals. Then based on these tests, the authors can assess the quality of their quasar+host decomposition and the significance of the host detection. If the range of results (as quantified by the defined metrics) for the quasar fits with all PSF stars is similar to that for the tests of stars fitting stars, then there is not much evidence for host detection of the quasar. However, if the authors find significantly different results when fitting the quasar with all PSF stars, then they may have a case of real host detection.

Reply: We have performed the tests described by the referee and presented the residuals of PSF star - PSF star in Figure E2. None of them show any extended emission that might be mistaken for a host. Furthermore, to quantitatively compare the level of PSF mismatch with the significance of the host light, we show the signal-to-noise per pixel of the PSF-subtracted quasar image, where the noise (standard deviation) is measured directly from the PSF star subtraction simulations. The figure clearly shows that the extended host light is well above the noise due to the PSF mismatch level across many contiguous pixels.

More importantly, we manually checked the individual fitting results from each PSF. In general, they are consistent with each other. We agree that there may be a better way than the chi^2 value to evaluate the performance of a PSF. However, using a different way to combine these PSF's results would not change our final result beyond the uncertainty level reported in this work.

However, detecting the host emission does not mean that the host measurements must be accurate. This is a critical issue for reliable stellar mass estimates, as well as constraints on morphological properties of the host. My concern is that the systematic uncertainties of the host measurements are substantial, and currently under-appreciated in quasar host studies with JWST imaging. To fully address this issue, the best approach is to simulate NIRCam images of z=6 quasar+host systems with known input host parameters, and apply the PSF templates to quantify the accuracy of recovered host parameters (e.g., centroid, flux, size and Sersic index). This is a worthwhile exercise for this important paper on first detection of z~6 quasar hosts in star light.

Reply: Following the suggestion, we have performed the simulation test using realistic JWST data and PSFs to confirm the accuracy of our host recovery. The details are presented in the Methods section (page 27, first paragraph). In this test, we use a random PSF from the library, add a host galaxy and use the other PSFs from the PSF library to fit the simulation. Our results (see Figure E3) clearly demonstrate that our modeling approach is able to recover the host parameters accurately.

Through all these tests above, we are confident in our measurement of host properties.

* Minor suggestions

I will provide a more complete list of minor comments when this manuscript is in a more advanced stage. Below are some points that I currently have.

- In several places (including the summary paragraph) the authors claim that "there are no host detection in star light beyond z>2". I am not sure this is true. HST-WFC3 F160W is able to detect host starlight at z>2. There are papers on such measurements at z>2 (whether or not these host measurements are robust is a different question; but certainly some of them look better than the results in this paper), e.g.,

https://ui.adsabs.harvard.edu/abs/2015ApJ...806..218G/abstract, https://ui.adsabs.harvard.edu/abs/2022arXiv220911249C/abstract. If you don't rely on direct imaging, then SED fitting arguably can also derive host star light for z>2 quasars (e.g., https://ui.adsabs.harvard.edu/abs/2016ApJ...824...70D/abstract). So my suggestion is to simply modify this claim to refer to "z>4 quasars" across this paper.
Reply: We agree that our claim was not accurate. We have now changed the 'z>2' to 'z□3'. We also modified the sentence and cited the useful references mentioned by the referee.

- Somewhat related to my major concern above, I think most people working on quasar-host imaging decomposition would agree that it is very difficult to get the host morphology right in these quasar-light-dominated systems. I think the authors should emphasize this caveat more, and caution on the interpretation of some best-fit parameters.
Reply: We agree. We have modified the Methods section "Tests of the robustness of the host galaxy properties" in the current version to emphasize this point. Our new simulation tests in the "Tests of the robustness of the host galaxy properties" section also verify the accuracy of the morphology measurement based on our decomposition approach.

- I am particularly skeptical about the claimed centroid offset of the hosts, unless the authors can demonstrate the robustness of the centroid constraints with simulated images + observed PSFs. I am also confused about the claim ".. similar offsets can be found in ~20% of lower redshift quasars [Ref 13,24 by the same primary authors]" with the same concern on the robustness of host centroid measurements. I would like to see more discussions on this, and citations to independent works supporting or against these low-z results.
Reply: As described in the Methods section "Tests of the robustness of the host galaxy properties" (paragraph #3), we have carried out simulations by inserting model Sersic profiles with different unresolved model quasar emissions based on the individual empirical PSFs. We then run the decomposition and check the offset between the host and the quasar. The offsets as reported in this paragraph, are lower than measured for both quasars.

Using catalogs at our disposal (HSC (z<1)-Ref#15 and HST/WFC3 (z~1.5)-Ref.#13), we looked at the distribution of offsets to compare with those seen in our quasar sample. We do find that offsets larger than those seen here occur for ~20% of the sample. This may not be surprising since this fraction is likely to be in a merger which can cause such offsets. We realized that comparing the offsets between different data sets is an entirely new avenue of analysis which is better placed in a subsequent study. Even so, we think the offsets are worth mentioning. Now, we drop the 20% remark and simply point out that while these offsets are similar to those in other samples, the magnitude of the offsets is larger relative to the sizes of their hosts. We think this is an interesting observation to report. Unfortunately, we cannot find published studies on this topic to make fair comparisons.

- Fig. 3 lower panel: no error bars on the data points?
Reply: Yes, the error bars were missing in that figure. However, as mentioned, the previous Figure 3 has now been removed due to the word limit.

- Table 1: rows on positional offset (") and (kpc). I do not understand why the fractional errors (%) would be different between the angular and physical offsets.

Reply: Thank you for pointing this out. We have corrected this in the current revision.

- Methods. About background subtraction, how were the close neighbors to J2255 subtracted? Or were the two close neighbors part of the decomposition fitting?
Reply: We modeled J2255 and its closest neighbor (object #1) simultaneously. We made this point clear in the Methods section “2D image decomposition of quasar and host galaxy emission using Galight”.

- About the final fitting results with uncertainties. In Methods it states “ We take the results from the five top-performing PSF models in the updated library. We determine our final result parameters with uncertainties by weighting their chi2 values.” How exactly is this done? E.g., is the best-fit parameter the 1/chi2-weighted mean of the five models? What about the uncertainties? Are the uncertainties marginalized?
Reply: We now have made clear citations to the equations we used from the reference.

- Methods on “Stellar mass and luminosity of the host galaxy”. NIRSpec data are mentioned here to justify the neglecting of line emission in the broad-band photometry. This is another (minor) reason why these NIRSpec data should be published along with this paper.
Reply: The NIRSpec data are now included and used to quantify the emission line contribution to the extended emission we detect in the NIRCam F356W images. We address this point in our reply to the first referee.

- A discussion on the systematic uncertainties on the stellar mass estimates from SED modeling would be useful.
Reply: To better estimate the systematics, we have improved our fit by releasing the SED parameters as free. Further discussion of the systematic uncertainties is included in the current version of the papers. This point is also related to our responses to the first referee’s comments.

We summarize other updates on the text as follows:

- The NIRSpec data is presented in Figure 3 to be quoted in the main text. The emission line measurements and the estimated Hbeta-based virial BH masses are reported in Table E2. The location of the two quasars in the BH mass – stellar mass plane and our constraint on the underlying co-evolutionary relation at z=6.4 are presented in Figure 4 and E5. The details of the spectroscopic observations, data analysis, and the mock observation of z=6.4 quasars are presented in the Methods sections.
- After performing the NIRCam simulation tests, we found that the F150W images give more precise host decomposition results if we drizzle the pixel scale to improve the pixel sampling by a factor of two. In the current version of the paper, we re-run the data reduction and perform the analysis. The improved inferred host values and systematics are included in Table E1.
- Given that the host galaxy of J2236 is detected in both bands, we now refer to it as the “first target”. As mentioned, due to page limitations, we have removed the size vs luminosity studies, to be presented in a future paper. The figures and tables are updated correspondingly.
- We updated the writing, figures, references, and the order in the author list wherever necessary. These minor corrections are highlighted in the text.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

Dear Editor,

I apologise for my response.

I am grateful to the authors for their additional tests to respond to my queries.

Regarding the contribution of the nebular lines to the photometry, it would be important that the authors report the observed line fluxes in a table.
On the extraction of the extended nebular component in J2255+0251, the authors report that they remove the nuclear component by using the 2D NIRSpec spectrum and “fit a Gaussian to the unresolved spatial profile at each wavelength and centroid to vary linearly with wavelength and letting the amplitude be a free parameter”; while the amplitude can be a free parameter, the width of the “unresolved” profile must be fixed to the JWST PSF at each wavelength, have the authors done so? Leaving the width free to vary may allow the extended component to be incorporated in the putative “unresolved” component.
Anyway, I am not sure I understand this test is conclusive in any way, as in the case of J2255 the slit is not aligned along the putative host galaxy major axis.

The determination of the stellar mass, and of the other stellar population parameters, based on only two photometric points (one of which being only an upper limit for one of the two galaxies), I believe is still a major problem of the paper.
Two single photometric points can be fitted with an infinity of stellar population models.
I hope there is agreement on the fact that, in science, with only two data points it is possible to constrain at maximum only two independent parameters.
In the paper, with only two data points the authors are constraining more than 5 independent variables: mass, metallicity, age, extinction and ionization parameters (and certainly the star formation history is also another embedded additional set of parameters, although it is not clear what the authors assume for the SFH).
I hope there is agreement that, mathematically, two data points cannot constrain 5 independent variables.
I believe that the impressive accuracy (0.2 dex) with which the authors derive the stellar mass of the host galaxies is simply a consequence of their tight priors on the various quantities. Indeed, in Fig. E4 there is a clear correlation between stellar mass and extinction and, to a lesser extent, between stellar mass and age (although difficult to assess as they explore a narrow range of ages). I’m sure that if they allow a larger range of extinction and ages then they would get a broader distribution of masses. One should keep in mind that large levels of extinction in these massive primeval galaxies is often found at high redshift (also based on the ALMA measurements giving large amount of dust masses).

Therefore, even if they have demonstrated that the extended emission is tracing the stellar light of the host galaxy (modulo my queries above), I do not think that the inferred stellar masseses are reliable.

If the result is deprived of the stellar mass result (and, therefore, of the M_bh/M_star result), the paper can still potentially be published with the other potentially more solid results, such as the disk-like morphology of the host galaxy and the fact that the quasar is displaced relative to the galactic centre. I am not totally sure if these two results alone warrant publication on Nature... potentially yes, but really borderline.

Referee #2 (Remarks to the Author):

I appreciate the authors' efforts of adding the NIRSpec data and performing additional simulations to test the robustness of host recovery. In my opinion, this version is much stronger and exciting than the previous version. I have a few remaining comments for the authors to address. After that I am happy to recommend it for publication.

The most significant comment is that I am surprised that it is even possible to derive a stellar mass estimate from two-band photometry with only 0.2 dex uncertainty. Even for $z<1$ quasar hosts with decent HST imaging, the stellar mass uncertainties from two-band photometry are typically >0.2 dex. I will elaborate below, but I think the authors need stronger arguments to support these rather precise M* estimates, or acknowledge that the actual uncertainties are larger. This is a different question than detecting the host starlight, which the authors have convinced me.

In the response, the authors mentioned that they re-drizzled the F150W images, which led to significant changes in the derived host mag in F150W for J2236 compared with the quoted mag uncertainties. This worries my that the realistic uncertainty in the host mag may be larger than the quoted formal errors. In fact, the estimated recovered mag rms in the simulations shown in Fig. E3 is also larger than all the quoted values in Table E1. I believe if a more realistic clumpy host galaxy (like the ones shown in Fig. E2 bottom) is used as input, the recovered host mags would have even larger rms scatter. Shouldn't the authors use these simulations to guide their estimation of the systematic uncertainty of recovered host mags? BTW, the F356W host mag+uncertainty for J2236 also changed slightly from the previous version, which lacks an explanation in the response.

But even with the <0.1 mag uncertainties in the derived host magnitudes, it seems implausible to constrain the stellar mass to better than 0.2 dex accuracy (0.15 dex from the fits) with only 2-band photometry. I suspect that the SED fitting model is too restrictive. It would be useful to discuss what realistic stellar mass uncertainties can be achieved, either with some simulations that cover a range of stellar population properties (with varying IMFs) or using a different SED fitting tool/methodology.

MINOR POINTS

I think the discussion on the potential spatial offset between the quasar and host galaxy centroid is still highly speculative. The authors themselves noted that the offset can be explained by dust obscuration or host structural inhomogeneities (or perhaps residual PSF mismatches). I strongly suggest that the authors dial it down on exotic explanations, which not only are not substantiated, but also creates distraction from the most important results of this paper. In general it's OK to mention it briefly in the text, but I'd suggest to remove it from the abstract.

In the main text, it was mentioned that the Sersic index is fixed to 1 for J2236. It would be useful to mention that relaxing the Sersic index value in the fit produces n=1.55 (so that the claimed “disk” morphology of the host is not due to the fixed n in the fit).

Author Rebuttals to First Revision:

Let me summarize the views of the referees. Neither referee thinks that the stellar mass is reliable. Referee 1 points out that the paper then rests on the disk-like morphology and offset. Referee 2 thinks that the discussion of the offset is very speculative. Putting on my astronomer's hat, I have stared at Figure 2 (data - point source) and see evidence for believable extended (and disk-like) emission only for F356W for J2236+0032. Although this is insufficient for Nature, Nature Astronomy may well be interested in a further revised paper, so I have included a link to make the transfer.

We thank you for the reconsideration of our manuscript.

Referees' comments:
Referee #1 (Remarks to the Author):

Regarding the contribution of the nebular lines to the photometry, it would be important that the authors report the observed line fluxes in a table.
Reply: We have reported the line fluxes in table E2 in this version.

On the extraction of the extended nebular component in J2255+0251, the authors report that they remove the nuclear component by using the 2D NIRSpec spectrum and "fit a Gaussian to the unresolved spatial profile at each wavelength and centroid to vary linearly with wavelength and letting the amplitude be a free parameter"; while the amplitude can be a free parameter, the width of the "unresolved" profile must be fixed to the JWST PSF at each wavelength, have the authors done so? Leaving the width free to vary may allow the extended component to be incorporated in the putative “unresolved” component.

Reply: As the referee suggested, we fixed the width of the unresolved profile to match the JWST PSF at each wavelength. We have modified the text to make it clear.

Anyway, I am not sure I understand this test is conclusive in any way, as in the case of J2255 the slit is not aligned along the putative host galaxy major axis.

Reply: We agree that the slit is not aligned with the host’s major axis, which can be considered a limitation of this measurement. However, the fact that the host galaxy is extended with a minor axis broader than the scale of the PSF (as shown in Figure 2) allows us to detect the extended emission from the spectrum. We have clarified this in the text.

The determination of the stellar mass, and of the other stellar population parameters, based on only two photometric points (one of which being only an upper limit for one of the two galaxies), I believe is still a major problem of the paper.Two single photometric points can be fitted with an infinity of stellar population models. I hope there is agreement on the fact that, in science, with only two data points it is possible to constrain at maximum only two independent parameters. In the paper, with only two data points the authors are constraining more than 5 independent variables: mass, metallicity, age, extinction and ionization parameters (and certainly the star formation history is also another embedded additional set of parameters, although it is not clear what the authors assume for the SFH). I hope there is agreement that, mathematically, two data points cannot constrain 5 independent variables.

Reply: We have added more extensive SED fitting tests and further discussion to demonstrate that a reliable stellar mass can be achieved with two carefully chosen photometric bands. For our targets at z~6.4, the filter selection of F150W and F356W was specifically designed to bracket the 4000A region and cover a broad wavelength range, thereby providing sensitivity to changes in stellar templates, essentially accounting for the strength of the young and old stellar population with insensitivity to changes in reddening which offset those of stellar age (as demonstrated in Bell & deJong 2001). In other words, the input parameters are not independent variables but rather correlated which leads to constrained masses. Even so, we would like to clarify that our intention was not to constrain five or more parameters with only two data points. Rather, we aimed to estimate the stellar mass of our host galaxy, which is known to be primarily sensitive to its luminosity longward of 4000A. While two photometric points can result in an infinite number of stellar population models, the overall amplitude of these models can still be constrained, given the long baseline of the two filters. It is important to note that our M* value is directly related to these amplitudes. As explained below, we have included additional caveats to the systematic uncertainties in the M* estimates.

It is worth stressing that the design choice (i.e., filter combination) was instrumental in the approval of our JWST proposal. We conducted SED simulation tests to support this claim, now described on page 27 of the current manuscript. The results show that the inferred M* value has a random scatter of 0.21 dex for J2236, consistent with the value obtained from the MCMC inference in Figure E4. If we had used the F150W+F200W filters (both below 4000A), the scatter would have increased to 0.61 dex. We also performed a test using a four-band filter (F150W, F200W, F277W, and F356W), resulting in a somewhat improved scatter of 0.17 dex.

Please keep in mind that we also have accurate distances to our targets based on their spectroscopic redshifts; unlike many inferences of stellar mass from photometry in the literature, we are not using the photometry to also estimate redshift. We also have a firm upper limit of 0.8 Gyr for the age of the stellar population, given the age of the universe at redshift 6.4.

I believe that the impressive accuracy (0.2 dex) with which the authors derive the stellar mass of the host galaxies is simply a consequence of their tight priors on the various quantities. Indeed, in Fig. E4 there is a clear correlation between stellar mass and extinction and, to a lesser extent, between stellar mass and age (although difficult to assess as they explore a narrow range of ages). I'm sure that if they allow a larger range of extinction and ages then they would get a broader distribution of masses. One should keep in mind that large levels of extinction in these massive primeval galaxies is often found at high redshift (also based on the ALMA measurements giving large amount of dust masses).

Reply: The impressive M* accuracy can be a consequence of the tight prior. However, the high-redshift nature of our host galaxies allows us to set this prior range reliably. For example, the prior stellar age covers the entire cosmic age up to z~6.4. The prior value of the extinction (Av) is also consistent with most studies in the current literature. We now mention in the paper that, for J2255, the detection of the narrow emission lines by NIRSpec allows us to perform a Balmer decrement measurement, from which we get an Av = 0.67±0.13 (Onoue et al. in prep). For J2236, we adopted a wider Av range as [0.3-2] to ensure the final inferred M* uncertainty by MCMC inference covers the systematic error.

Therefore, even if they have demonstrated that the extended emission is tracing the stellar light of the host galaxy (modulo my queries above), I do not think that the inferred stellar masses are reliable.If the result is deprived of the stellar mass result (and, therefore, of the M_bh/M_star result), the paper can still potentially be published with the other potentially more solid results, such as the disk-like morphology of the host galaxy and the fact that the quasar is displaced relative to the galactic centre. I am not totally sure if these two results alone warrant publication on Nature... potentially yes, but really borderline.

Reply: We hope our updated tests and discussions will convince the referee that the reliability of our M* is sufficient to compare with the single-epoch black hole masses (with 0.4 dex systemic error). We have added the following line to the paper to recognize the likely uncertainty in the stellar mass measurements, "*We use simulations and three different SED fitting codes to demonstrate that the results are robust, although further* *observations will lessen potential systematic errors in these mass estimates*".

Referee #2 (Remarks to the Author):

I appreciate the authors' efforts of adding the NIRSpec data and performing additional simulations to test the robustness of host recovery. In my opinion, this version is much stronger and exciting than the previous version. I have a few remaining comments for the authors to address. After that I am happy to recommend it for publication.

Reply: We thank the referee again for the comments which have improved the paper (see below) and his/her recommendation for publication after this revision.

The most significant comment is that I am surprised that it is even possible to derive a stellar mass estimate from two-band photometry with only 0.2 dex uncertainty. Even for z<1 quasar hosts with decent HST imaging, the stellar mass uncertainties from two-band photometry are typically >0.2 dex. I will elaborate below, but I think the authors need stronger arguments to support these rather precise M* estimates, or acknowledge that the actual uncertainties are larger. This is a different question than detecting the host starlight, which the authors have convinced me.

Reply: In this version, more effort has been spent on testing the systematic uncertainties of estimating stellar masses based on two-band photometry. We now use three different SED codes with a wider prior range to infer the probability distribution of M*. The final uncertainty now is ~0.23 dex for J2236, and ~0.28 dex for J2255. Most of the updates are summarized in the SED fitting section in the Methods section.

As mentioned by the referee, there are some works in the literature that use two-band photometry by HST to constrain the stellar mass of quasar hosts at z < 1 with larger uncertainty. However, the high redshift nature of our host galaxies allows a tighter prior (e.g., stellar age < 0.8 Gyr, i.e., the cosmic age). Thus, it is not surprising for us to get stellar mass estimates with higher precision than those z<1 galaxies.

In the response, the authors mentioned that they re-drizzled the F150W images, which led to significant changes in the derived host mag in F150W for J2236 compared with the quoted mag uncertainties. This worries my that the realistic uncertainty in the host mag may be larger than the quoted formal errors. In fact, the estimated recovered mag rms in the simulations shown in Fig. E3 is also larger than all the quoted values in Table E1. I believe if a more realistic clumpy host galaxy (like the ones shown in Fig. E2 bottom) is used as input, the recovered host mags would have even larger rms scatter. Shouldn't the authors use these simulations to guide their estimation of the systematic uncertainty of recovered host mags?

Reply: We are grateful for the caveats raised by the referee. We now have used the simulation result (i.e., 0.15 mag scatter) to guide our SED fitting. Combining with the current JWST calibration uncertainty (<0.1 mag), we have adopted a 0.20 mag uncertainty in the SED fitting in the current version. We have clarified this in the paper.

BTW, the F356W host mag+uncertainty for J2236 also changed slightly from the previous version, which lacks an explanation in the response.

Reply: We apologize for overlooking this change in our previous reply report. In the first version draft, the fitting result of F2236 by the averaged PSF was not applied properly. We had noticed and solved this issue, and the mag+uncertainty is updated slightly for J2236.

But even with the <0.1 mag uncertainties in the derived host magnitudes, it seems implausible to constrain the stellar mass to better than 0.2 dex accuracy (0.15 dex from the fits) with only 2-band photometry. I suspect that the SED fitting model is too restrictive. It would be useful to discuss what realistic stellar mass uncertainties can be achieved, either with some simulations that cover a range of stellar population properties (with varying IMFs) or using a different SED fitting tool/methodology.

Reply: We now have added more discussion regarding the SED fitting to the paper.

We also tested using different SED modeling codes (Bagpipes and CIGALE); together with gsf, they all yield consistent M* estimates. A SED fitting simulation is also performed to demonstrate the accuracy of stellar mass estimates as a function of the number of filters employed and wavelength coverage.

We understand that different IMF assumptions will lead to varying stellar mass estimates. We now have updated the draft in which we explain that we use a Charbier IMF to derive a consistent M* so that we can make a relatively fair comparison with the local scaling relations between black hole mass and galaxy mass.

MINOR POINTS

I think the discussion on the potential spatial offset between the quasar and host galaxy centroid is still highly speculative. The authors themselves noted that the offset can be explained by dust obscuration or host structural inhomogeneities (or perhaps residual PSF mismatches). I strongly suggest that the authors dial it down on exotic explanations, which not only are not substantiated, but also creates distraction from the most important results of this paper. In general it's OK to mention it briefly in the text, but I'd suggest to remove it from the abstract.

Reply: We agree that this is not a major scope of our work. In the current version, we have dialed it down in the main text and the abstract.

In the main text, it was mentioned that the Sersic index is fixed to 1 for J2236. It would be useful to mention that relaxing the Sersic index value in the fit produces n=1.55 (so that the claimed “disk” morphology of the host is not due to the fixed n in the fit).

Reply: We have now made updates and mentioned this in the main text.

Other minor changes:

- In the process of SED modeling for J2255, we made a modification to the settings of the F150W's flux upper limit by implementing a soft boundary. This change allows the SED templates more flexibility in terms of variations on the blue side of the spectrum. The same upper boundary is also set when testing the SED model using Bagpipes and CIGALE.
- Due to the updated M* estimates, the result and the figures related to the scaling relation are modified correspondingly.
- The label J2236+0032 has been corrected (from J2236+0036) in Figure 3.
- In Method, we removed the description of the photometry of objects #1 and #2 and Table E3 to shorten the length of the paper.
- We also improved the texts, and modified our language to shorten the paper.

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

Dear Editor,

I apologise for the delay in responding to this additional request.

I my previous report I had already expressed that the detection of the host galaxy in quasars at z~6, and the characterisation of their properties, is an interesting result which may be worth publishing on Nature (although AGN host galaxies at similar redshift have been detected at similar redshifts with JWST, although in lower luminosity AGN).

I think that the detection of the host galaxy stellar light is solid, although I am not sure about the small uncertainties quoted for the host galaxy photometries, given the large residuals in the modelling (reaching +- 6 sigma).

My main concern was on the analysis were about the stellar masses, based on only two photometric points.

I have re-estimated the stellar masses based on their photometries, by using two different SED fitting tools.

I have used the same assumptions as the authors, except for the following few aspects:

1) For setting the upper limit on the age uniform prior they use the “age of the Universe” of 750 Myr at z=6.4; however, if I'm not mistaken, the age of the universe at that redshift with their chosen cosmology is 840 Myr. So I have used this upper limit on the age prior.

2) Regarding the prior on the extinction, I do not agree that the Balmer decrement of the narrow components can provide a meaningful proxy for the extinction to the host galaxy. Leaving aside broad/narrow component decomposition problems, this Balmer decrement of the narrow components may provide the extinction towards the Narrow Line Region, which often can have a much lower extinction than the host galaxy of AGN.

So I have used a flat prior to much higher extinctions, Av between zero and 5.

3) Although, as discussed above, I am somewhat doubtful about the photometric uncertainties, I have used the value of +-0.2 mag that they give in the text. However, I note that in their table E1 they give much smaller uncertainties. So I prompt them to provide consistent uncertainties in the table.

With the modifications above, I do recover stellar masses similar to those obtained by the authors, although with somewhat larger uncertainties (~0.3-0.4 mag).
Given that this responds to my main concern, I am in favour of recommending publication on Nature, provided that the authors re-fit the photometric points by possibly complying with the points raised by me above and, therefore, re-assess the uncertainties on the stellar masses accordingly.

Referee #2 (Remarks to the Author):

I thank the authors for sending their NIRCam data on J2236. My group and I have analyzed these data, and we have looked carefully into the methodologies of image decomposition and SED fitting with NIRCam imaging. To summarize, we found some inconsistencies with the decomposition that mostly affected the F150W host measurements in J2236. There are additional concerns regarding PSF mismatches and the stellar mass estimates with SED fitting. These concerns are significant and impact the interpretation of the data.

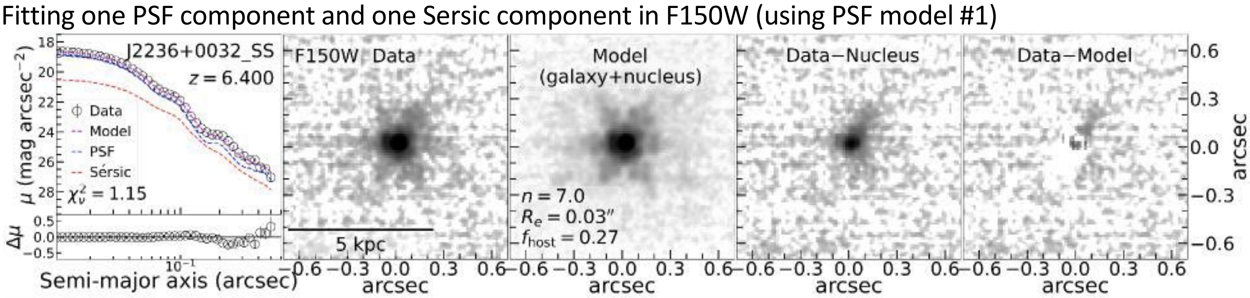
- Decomposition using the provided PSF models.

Using the same PSF models as provided, our decomposition results are slightly different in F356W but significantly different in F150W. These comparisons are detailed below.

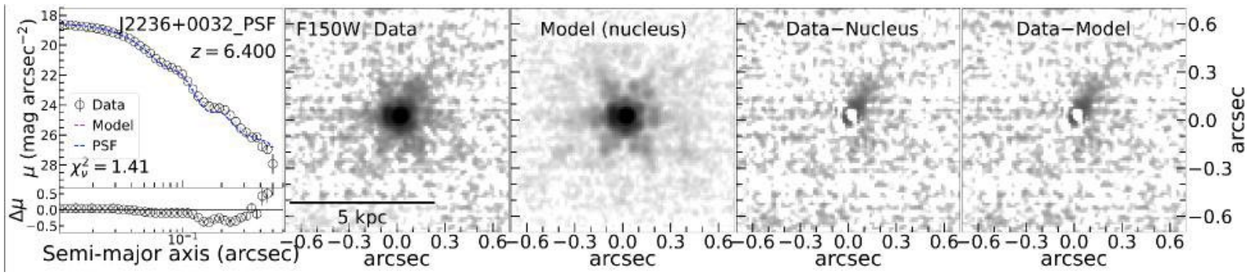
For F356W, the 5 PSF models result in host mag of 22.96-23.18, which is systematically brighter than 23.25+-0.10 derived by the authors. For F150W, the 5 PSF models (fixed n=1) result in host mag of 24.65-24.96, which is significantly brighter than 25.32+-0.08 derived by the authors. The detected "host flux" in F150W by the authors has little central contribution (Fig 2 in the paper), and is dominated by a feature to the upper right that apparently aligns with one PSF spike. Assuming it is real emission (not some weird PSF structures at the location of the quasar), it is largely spatially separate from the host component in F356W. Perhaps this F150W feature is a satellite galaxy rather than the host itself. But the stellar mass estimation assumes that they are the same stellar population in F150W and F356W.

In our decomposition, however, there is a central component of the host in addition to the upper-right plume feature (see figures below). Finally, using a single PSF model (no host component) for F150W produces a chi2/dof=1.41, which is still acceptable.

Fitting one PSF component and one Sersic component in F150W (using PSF model #1)



Fitting only a PSF (no host) in F150W (using PSF model #1)



The full table and figures of fitting results (5 PSF models in each band, with additional tests; the columns should be self explanatory) is attached in a zip file. The main issue is with the F150W decomposition. The F356W results are mostly fine, within the 0.2 mag inflated errors adopted by the authors. However, the differences from the authors' results are systematic, i.e., we are getting higher host fluxes in both bands. These tests suggest that even with the same PSF models, there can be systematic differences in the decomposition results from different fitting methodologies, which in the case of F150W are significantly larger than the quoted formal uncertainties. [See my next comment on mismatched PSF models.]

These systematic differences stem from different decomposition methodologies (or default settings in the fitting parameters). I strongly encourage the authors to use GALFIT to reproduce their GALIGHT results. The former tool has been thoroughly tested in the community, while the latter is a relatively new tool and is yet to be tested. For example, there are significant differences in the host photometry derived for the same quasar SDSS1420+5300A using NIRCam imaging analyzed by the same principal authors (arXiv:2209.03359, using GALIGHT) and by an independent study (arXiv:2304.13776, using GALFIT). Our investigations on this matter indicate that the host measurements in arXiv:2209.03359 with GALIGHT are biased.

- Issues with PSF and host model mismatches

These imaging decomposition and host measurements in z~6 quasars are particularly challenging, even for JWST. There is a recent study (arXiv:2304.13776) on JWST NIRCam PSF variations and AGN-host decomposition. Using mis-matched PSF models and/or inappropriate host components (e.g., 1 Sersic versus 2 Sersic) produces systematic uncertainties that can be much larger than the quoted uncertainties from the fit, especially for AGN-dominated systems.

In addition, a mismatched PSF does not necessarily lead to a larger chi2 compared with the correct PSF. So overall I am not convinced that the authors' approach of selecting the best PSF model is the right approach. Instead, the authors could quantify the systematic uncertainty in host photometry using the typical PSF variations in these JWST bands. Moreover, if the underlying host model is different from what's assumed in the fitting, then I'd expect additional uncertainties from assuming a single Sersic profile. Understandably a more complicated host model is not warranted given the data quality, but this caveat should be mentioned. Overall, these systematic uncertainties can add up to >0.2 mag even for bands where the host-to-AGN ratio is ~1. In F150W, the reported host fraction is <10%, where systematic biases from PSF and model mismatches can be much more severe.

- Stellar mass

Using the same host photometry derived by the authors and the same assumptions in the SED fitting, we derived a best-fit M*=10.74+-0.18 (with continuum Av prior=[0,0.82] mag, best-fit Av=0.3 mag), and M*=11.00+-0.22 (Av prior=[0,2.05], best-fit Av=1.3 mag). While these results are broadly consistent with the authors' result, there are some degeneracies. Given the bright NUV flux in F150W, one may argue that the low Av scenario might be more favorable. In general, the SED fits have more free parameters (4) than data points (2), and it is impossible to judge which model is physically correct (the fit will simply find the numerically best-fit model).

The stellar mass estimates are also severely affected by other assumptions (e.g., IMF) and the derived host photometry. In particular, the host F150W flux and the F356W flux derived by the authors have different spatial extent, and a joint SED fit to constrain the stellar population is questionable.

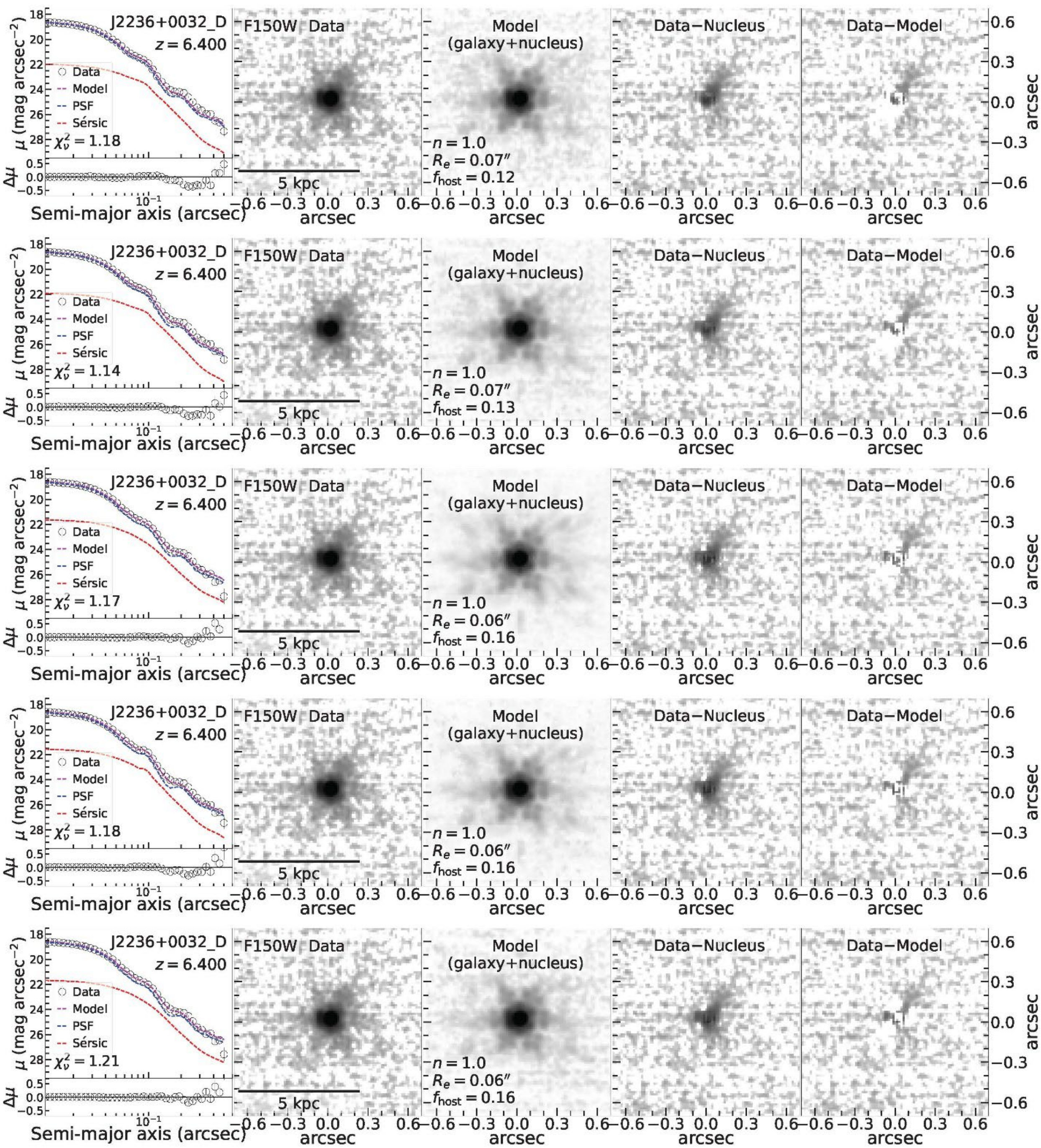
Overall I agree with the other referee that two-band photometry is insufficient to robustly constrain the stellar mass. Even with F356W you are only sampling rest-frame optical light, whereas the bulk of the stellar mass is traced in rest-frame NIR. In the case of J2255, there are only detection in one band and upper limits in the other. Given these data and caveats in image decomposition, I think it's practically impossible to constrain M* to be within 0.3 dex uncertainties, if you can constrain M* at all.

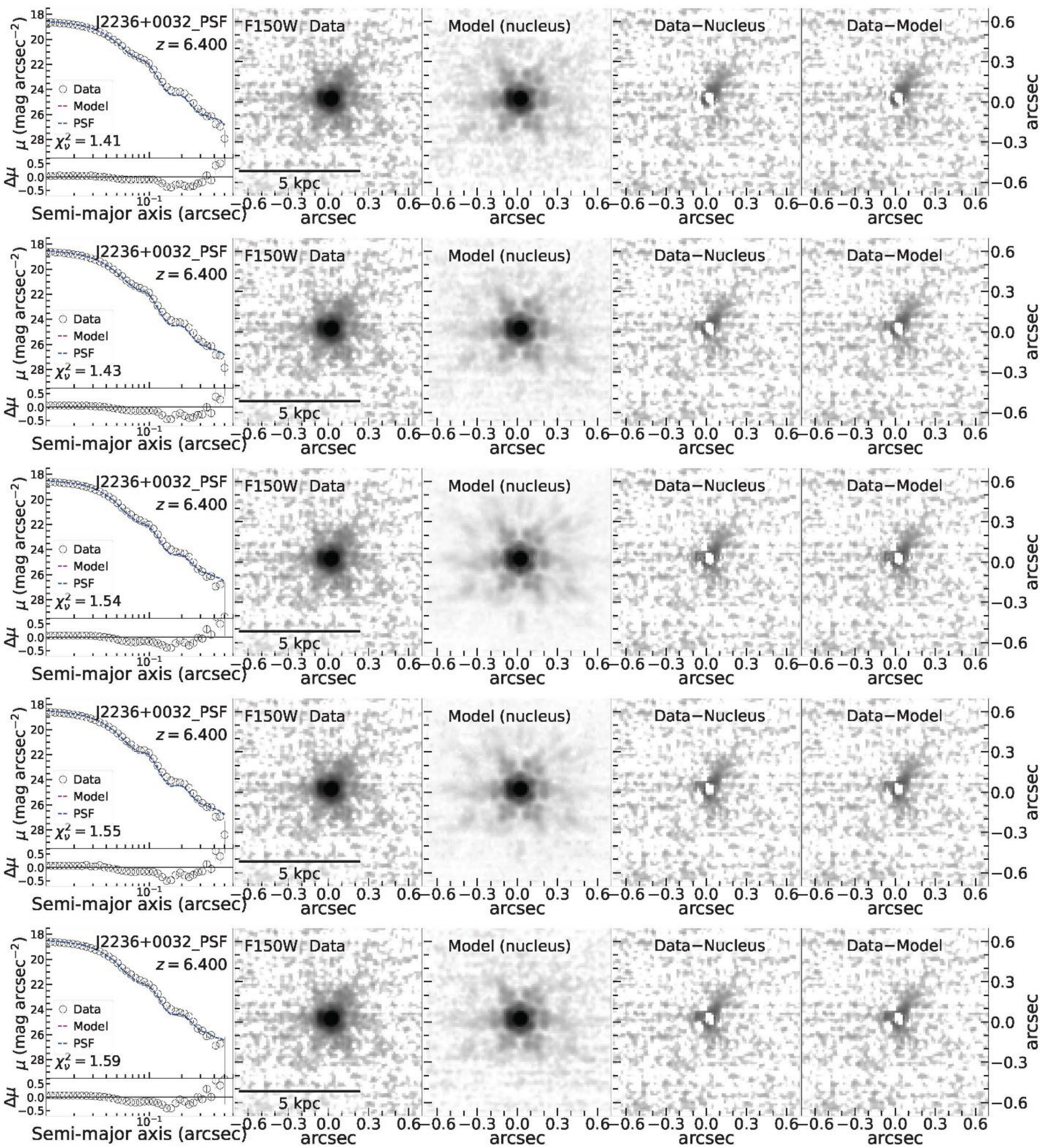
Given these considerations, my recommendations are the following.

- I think the F356W host detection is secure, and the stellar absorption features from NIRSpec further support the detection of the host stellar component. In my opinion, this result is important and I'd offer my support for a Nature publication.
- However, I have significant concerns that the NIRCam imaging analyses/results have substantial issues that will impact the repeatability of the results and interpretation of the data.
- The host measurement in F150W (for J2236) is questionable, as we were unable to reproduce the results. I urge the authors to investigate their GALIGHT fitting results, and compare with GALFIT results. Then redo the SED fitting part.
- The systematic uncertainties in the derived host photometry (as well as structural properties) from PSF and host model mismatches are substantially larger than the formal fitting uncertainties. I suggest that the authors make this point crystal clear, and
- add notes to Table E1 to caution on the reported formal fitting uncertainties. I also suggest the authors to cite this paper (arXiv:2304.13776) on NIRCam PSFs to highlight the caveats and challenges of measuring high-z quasar hosts with NIRCam imaging.
- The authors should explicitly state that the systematic uncertainties on the stellar mass estimates are significantly larger than their reported values. I am not convinced that

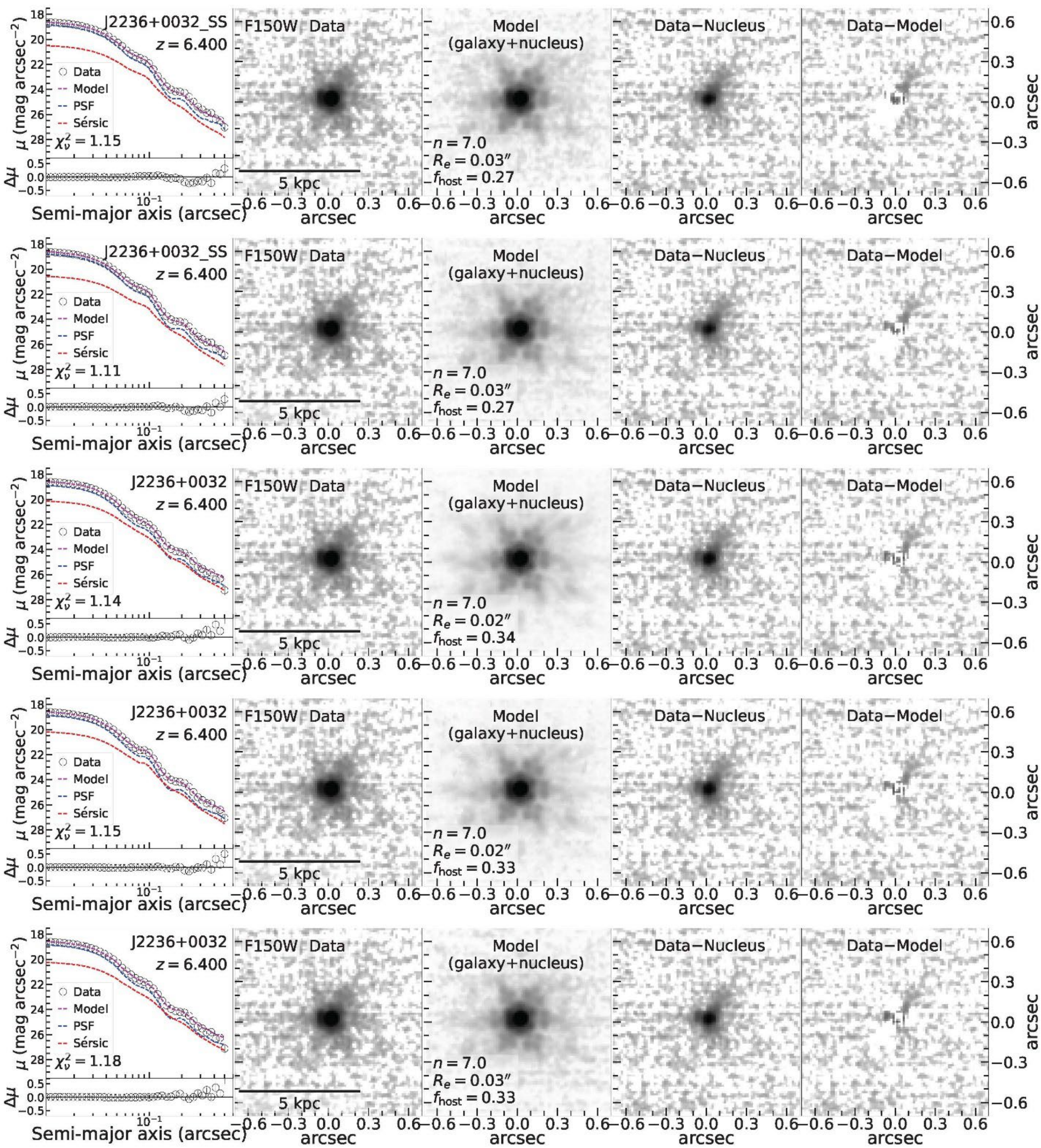
you can get <~0.3 dex uncertainty in M* with 1-2 data points, even if the host photometry is correct.

- The authors should remove the claim on offset AGN entirely from this paper. The data does not support such a claim, and the authors did not go to sufficient depth to rule out that this is due to decomposition systematics (e.g., fitting tools, mismatched PSFs, or systematic biases from systems with low host-to-AGN ratios). In particular, I point out that the most extreme host offset is in F150W for J2236, where the authors' decomposition lacks a central host component. I suspect that a similar bias exists for the J2255 decomposition.
- The strong stellar absorption lines from NIRSpec carry additional information on the stellar population. The authors should discuss this in the context of the NIRCam imaging results.

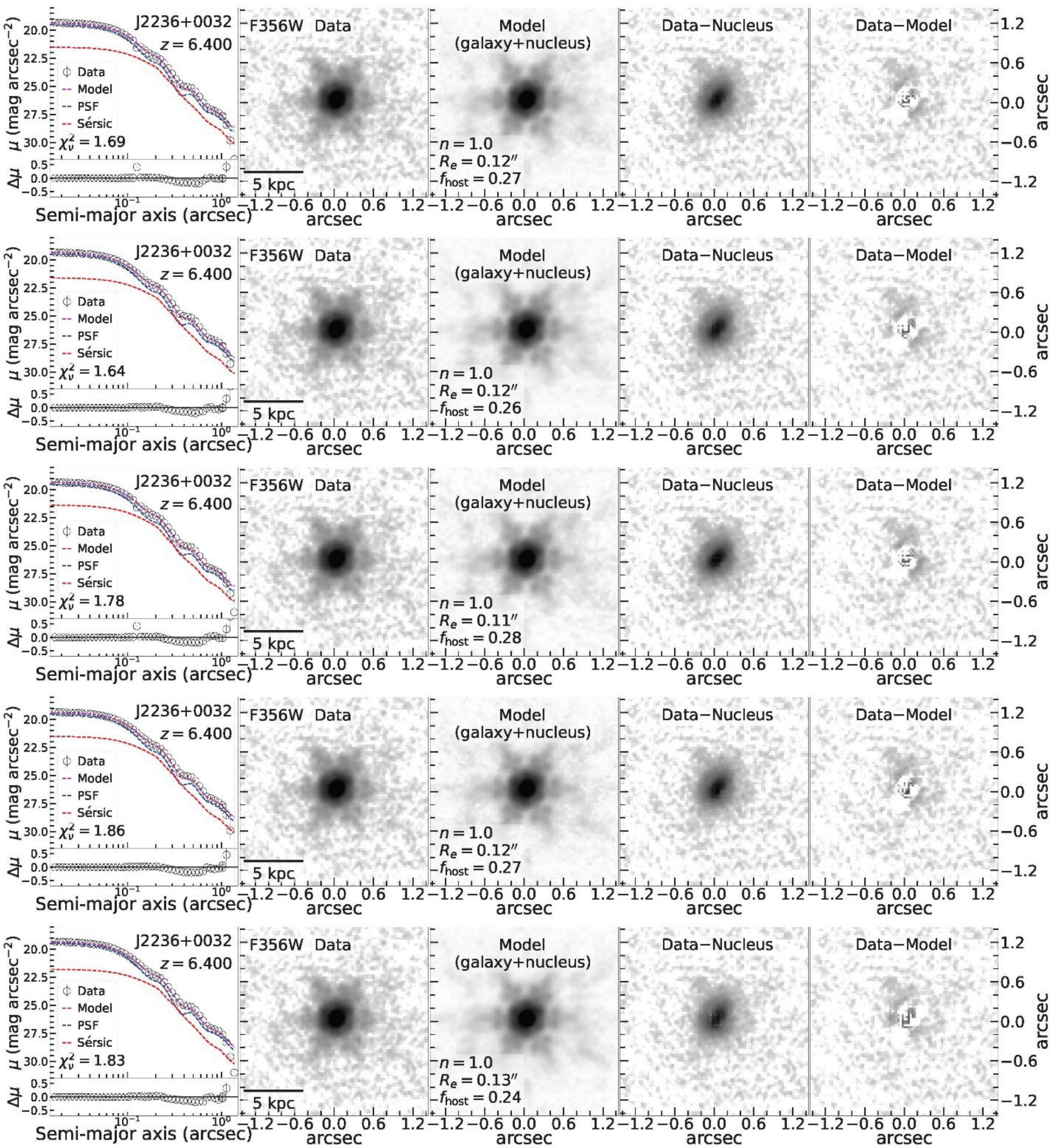

J2236+0032_D
$z = 6.400$
Data
Model
PSF
Sérsic
$\chi^2_\nu = 1.18$
μ (mag arcsec^{-2})
$\Delta\mu$
Semi-major axis (arcsec)
F150W Data
Model (galaxy+nucleus)
$n = 1.0$
$R_e = 0.07''$
$f_{\rm host} = 0.12$
5 kpc
Data−Nucleus
Data−Model
arcsec
J2236+0032_D
$z = 6.400$
$\chi^2_\nu = 1.14$
$n = 1.0$
$R_e = 0.07''$
$f_{\rm host} = 0.13$
J2236+0032_D
$z = 6.400$
$\chi^2_\nu = 1.17$
$n = 1.0$
$R_e = 0.06''$
$f_{\rm host} = 0.16$
J2236+0032_D
$z = 6.400$
$\chi^2_\nu = 1.18$
$n = 1.0$
$R_e = 0.06''$
$f_{\rm host} = 0.16$
J2236+0032_D
$z = 6.400$
$\chi^2_\nu = 1.21$
$n = 1.0$
$R_e = 0.06''$
$f_{\rm host} = 0.16$



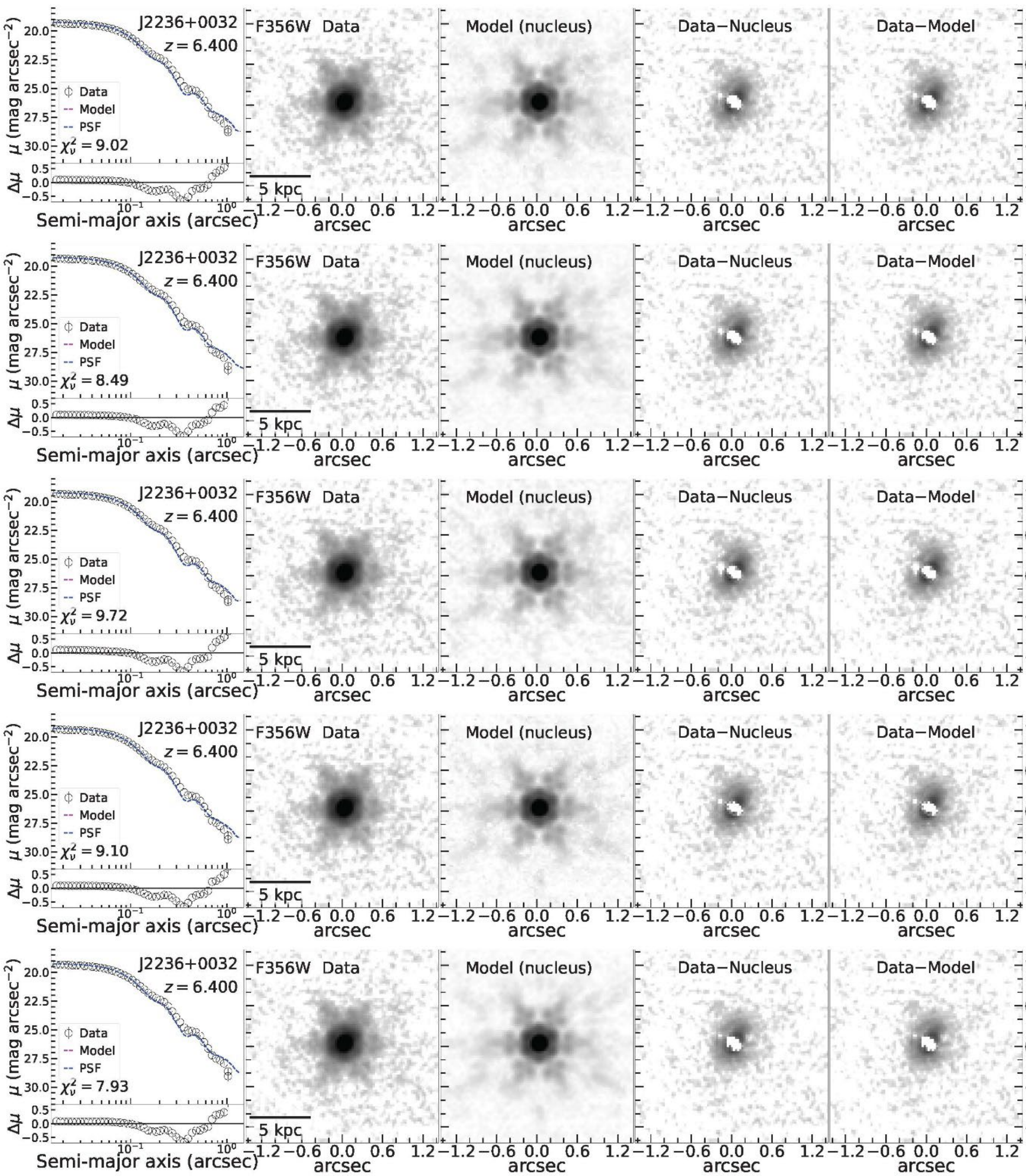
J2236+0032_PSF
z = 6.400
Data
Model
PSF
χ²_ν = 1.41
χ²_ν = 1.43
χ²_ν = 1.54
χ²_ν = 1.55
χ²_ν = 1.59
μ (mag arcsec⁻²)
Δμ
Semi-major axis (arcsec)
F150W Data
Model (nucleus)
Data−Nucleus
Data−Model
5 kpc
arcsec



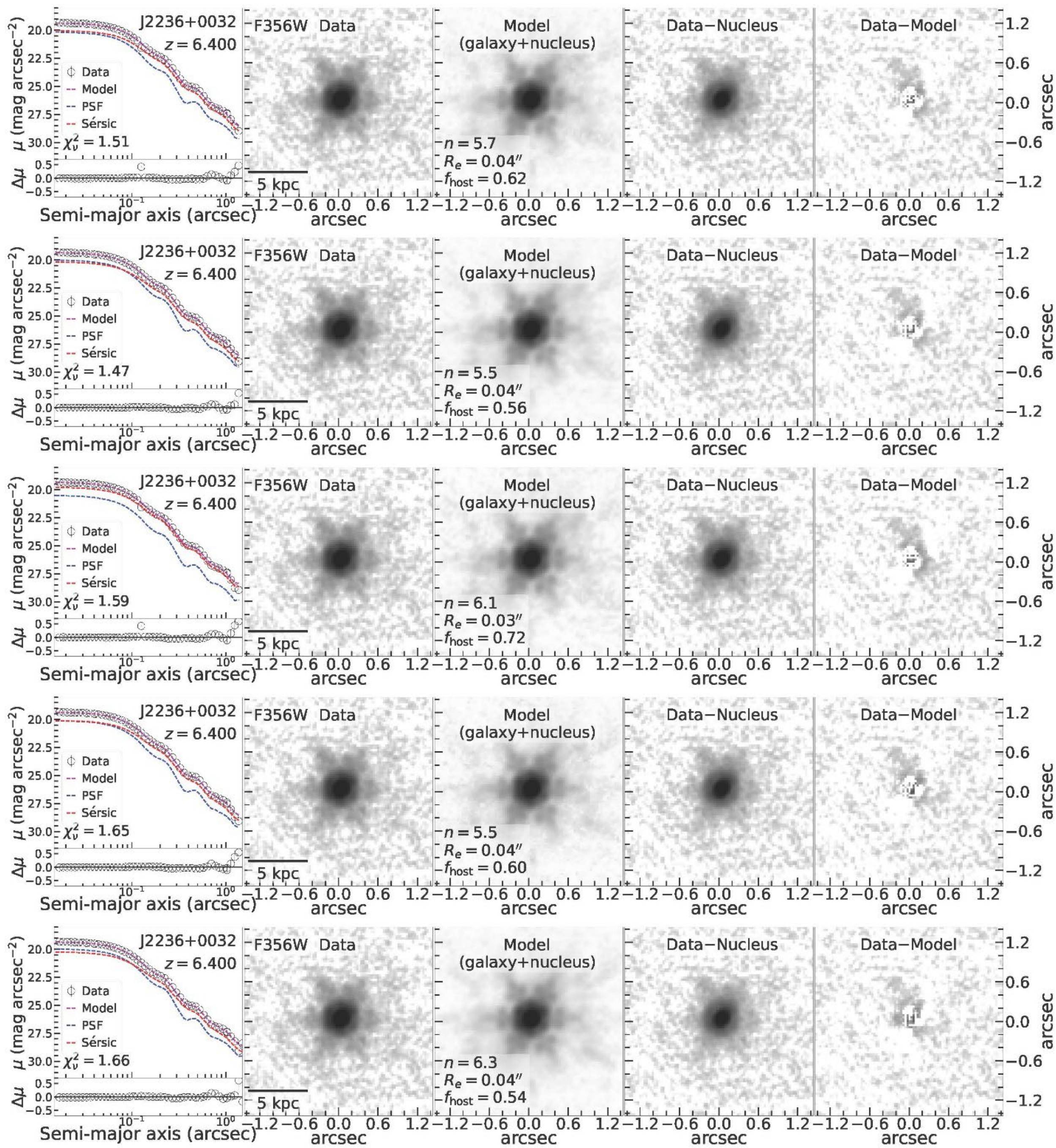
J2236+0032_SS
$z = 6.400$
Data
Model
PSF
Sérsic
$\chi^2_\nu = 1.15$
F150W Data
Model (galaxy+nucleus)
$n = 7.0$
$R_e = 0.03''$
$f_{\rm host} = 0.27$
5 kpc
Data−Nucleus
Data−Model
J2236+0032_SS
$z = 6.400$
$\chi^2_\nu = 1.11$
$n = 7.0$
$R_e = 0.03''$
$f_{\rm host} = 0.27$
J2236+0032
$z = 6.400$
$\chi^2_\nu = 1.14$
$n = 7.0$
$R_e = 0.02''$
$f_{\rm host} = 0.34$
J2236+0032
$z = 6.400$
$\chi^2_\nu = 1.15$
$n = 7.0$
$R_e = 0.02''$
$f_{\rm host} = 0.33$
J2236+0032
$z = 6.400$
$\chi^2_\nu = 1.18$
$n = 7.0$
$R_e = 0.03''$
$f_{\rm host} = 0.33$
μ (mag arcsec^{-2})
$\Delta\mu$
Semi-major axis (arcsec)
arcsec



J2236+0032
z = 6.400
Data
Model
PSF
Sérsic
χ²_ν = 1.69
Semi-major axis (arcsec)
μ (mag arcsec⁻²)
Δμ
F356W Data
5 kpc
Model (galaxy+nucleus)
n = 1.0
R_e = 0.12″
f_host = 0.27
Data−Nucleus
Data−Model
arcsec
χ²_ν = 1.64
R_e = 0.12″
f_host = 0.26
χ²_ν = 1.78
R_e = 0.11″
f_host = 0.28
χ²_ν = 1.86
R_e = 0.12″
f_host = 0.27
χ²_ν = 1.83
R_e = 0.13″
f_host = 0.24



J2236+0032
$z = 6.400$
Data
Model
PSF
$\chi^2_\nu = 9.02$
μ (mag arcsec^{-2})
$\Delta\mu$
Semi-major axis (arcsec)
F356W Data
Model (nucleus)
Data−Nucleus
Data−Model
5 kpc
arcsec
J2236+0032
$z = 6.400$
Data
Model
PSF
$\chi^2_\nu = 8.49$
F356W Data
Model (nucleus)
Data−Nucleus
Data−Model
5 kpc
J2236+0032
$z = 6.400$
Data
Model
PSF
$\chi^2_\nu = 9.72$
F356W Data
Model (nucleus)
Data−Nucleus
Data−Model
5 kpc
J2236+0032
$z = 6.400$
Data
Model
PSF
$\chi^2_\nu = 9.10$
F356W Data
Model (nucleus)
Data−Nucleus
Data−Model
5 kpc
J2236+0032
$z = 6.400$
Data
Model
PSF
$\chi^2_\nu = 7.93$
F356W Data
Model (nucleus)
Data−Nucleus
Data−Model
5 kpc



J2236+0032
z = 6.400
Data
Model
PSF
Sérsic
χ²ν = 1.51
μ (mag arcsec⁻²)
Δμ
Semi-major axis (arcsec)
F356W Data
Model (galaxy+nucleus)
n = 5.7
Re = 0.04″
fhost = 0.62
5 kpc
Data−Nucleus
Data−Model
arcsec
J2236+0032
z = 6.400
χ²ν = 1.47
n = 5.5
Re = 0.04″
fhost = 0.56
J2236+0032
z = 6.400
χ²ν = 1.59
n = 6.1
Re = 0.03″
fhost = 0.72
J2236+0032
z = 6.400
χ²ν = 1.65
n = 5.5
Re = 0.04″
fhost = 0.60
J2236+0032
z = 6.400
χ²ν = 1.66
n = 6.3
Re = 0.04″
fhost = 0.54

Author Rebuttals to Second Revision:

Replies to the referees:

Referee #1 (Remarks to the Author):

I have re-estimated the stellar masses based on their photometries, by using two different SED fitting tools.

I have used the same assumptions as the authors, except for the following few aspects:

1) For setting the upper limit on the age uniform prior they use the "age of the Universe" of 750 Myr at z=6.4; however, if I'm not mistaken, the age of the universe at that redshift with their chosen cosmology is 840 Myr. So I have used this upper limit on the age prior.

2) Regarding the prior on the extinction, I do not agree that the Balmer decrement of the narrow components can provide a meaningful proxy for the extinction to the host galaxy. Leaving aside broad/narrow component decomposition problems, this Balmer decrement of the narrow components may provide the extinction towards the Narrow Line Region, which often can have a much lower extinction than the host galaxy of AGN.

So I have used a flat prior to much higher extinctions, Av between zero and 5.

3) Although, as discussed above, I am somewhat doubtful about the photometric uncertainties, I have used the value of +-0.2 mag that they give in the text. However, I note that in their table E1 they give much smaller uncertainties. So I prompt them to provide consistent uncertainties in the table.

With the modifications above, I do recover stellar masses similar to those obtained by the authors, although with somewhat larger uncertainties (~0.3-0.4 mag).

Given that this responds to my main concern, I am in favour of recommending publication on Nature, provided that the authors re-fit the photometric points by possibly complying with the points raised by me above and, therefore, re-assess the uncertainties on the stellar masses accordingly.

We have now re-fitted our photometry to derive M* using the same priors as suggested by the referee and the final M* uncertainty (0.34 and 0.44 dex for J2236 and J2255, respectively) is consistent with the referee's results. We have also softened our reference to the Balmer decrement. Note that for J2255, we assumed a log-normal distribution on the extinction, following the JWST study of high-redshift galaxies by Shapley et al. 2023 arXiv:2301.03241. We thank the referee for confirming that the stellar masses we estimated are recovered.

Referee #2 (Remarks to the Author):

- I think the F356W host detection is secure, and the stellar absorption features from NIRSpec further support the detection of the host stellar component. In my opinion, this result is important and I'd offer my support for a Nature publication.

We thank our referee once again for the suggestions and his/her recommendation for publication.

- However, I have significant concerns that the NIRCam imaging analyses/results have substantial issues that will impact the repeatability of the results and interpretation of the data.

We have carefully tested the mismatch between our measurement and the ones obtained by the referee using Galfit, as described below.

- The host measurement in F150W (for J2236) is questionable, as we were unable to reproduce the results. I urge the authors to investigate their GALIGHT fitting results, and compare with GALFIT results. Then redo the SED fitting part.

We have used Galfit to fit our images and compared them with our Galight results. We realized that the mismatch mentioned by the referee is caused by a different PSF super-sampling parameter in Galight. As now described in our paper, this parameter controls the level of PSF interpolation that is used to shift the position of the PSF to fit the image of the quasar. In our previous result, the host results do not apply any PSF interpolation (super_sampling = 1).

Our new analysis shows that the Galight results are consistent with Galfit when super_sampling is set to 2. In our updated draft, our host properties are given as the average value using supersampling factors for both 1 and 2 with uncertainties spanning the difference between results based on these two settings. We also provide the Galfit results in Table E1. The Galfit and Galight host magnitudes and uncertainties are consistent with each other and with the independent check that , referee #2 carried out. To emphasize this point, we added a new figure E4, showing that the Galight result is in remarkable agreement with Galfit with PSF supersampling factor =2 using the same PSF for the decomposition with both codes.

We note that the new decomposition for J2236 using super_sampling = 2 nicely recovers the central flux in the F150W band as seen by the referee and his/her group. We greatly appreciate this effort which improved our analysis.

We also re-performed the SED fitting based on our new photometric measurements to derive stellar masses with more appropriate uncertainties.

- The systematic uncertainties in the derived host photometry (as well as structural properties) from PSF and host model mismatches are substantially larger than the formal fitting uncertainties. I suggest that the authors make this point crystal clear, and add notes to Table E1 to caution on the reported formal fitting uncertainties.

I also suggest the authors to cite this paper (arXiv:2304.13776) on NIRCam PSFs to highlight the caveats and challenges of measuring high-z quasar hosts with NIRCam imaging. The authors should explicitly state that the systematic uncertainties on the stellar mass estimates are significantly larger than their reported values. I am not convinced that you can get <~0.3 dex uncertainty in M* with 1-2 data points, even if the host photometry is correct.

As mentioned above, we have re-evaluated the systematic uncertainties of the host photometry. The paper mentioned by the referee is now prominently cited in the current version of our paper. After remeasuring our host photometry and applying the new SED prior, the M* uncertainties are 0.34 dex and 0.44 dex for J2236 and J2255, respectively, matching the forecast by our referees.

- The authors should remove the claim on offset AGN entirely from this paper. The data does not support such a claim, and the authors did not go to sufficient depth to rule out that this is due to decomposition systematics (e.g., fitting tools, mismatched PSFs, or systematic biases from systems with low host-to-AGN ratios). In particular, I point out that the most extreme host offset is in F150W for J2236, where the authors' decomposition lacks a central host component. I suspect that a similar bias exists for the J2255 decomposition.

We agree with the referee that the positional offset values may be related to the different PSF supersampling parameters, so more work would be needed to confirm the robustness of the offset between the quasar and the host galaxy. which requires more simulation tests and more investigations to confirm this detection. Doing so would go beyond the scope of the paper, so we have removed the claim on the offset entirely from this paper and will return to this question in future papers.

- The strong stellar absorption lines from NIRSpec carry additional information on the stellar population. The authors should discuss this in the context of the NIRCam imaging results.

We have added more discussion and presented more information on the stellar population. We also mentioned that more detailed studies will be presented in Onoue et al. in prep, which will focus on the NIRSpec data analysis.

Other minor changes:

- We have made updates to the flux correction for the quasar's 5100 Å continuum luminosity based on the new quasar decomposition result. This has resulted in slight changes to the estimated MBH mass. The revised values have been colored in blue in the draft, and Table E2 has also been updated accordingly.

- The paper is purely focused on host galaxy and quasar mass measurements. The discussion on a possible offset of the host galaxy from the quasar are removed. Hence, the R_1/2 and positional offsets are also removed in Table E1.

- The paper and figure format has been updated to align with the Nature formatting guidelines. We also use new LaTex template to compile our manuscripts.