We thank the reviewers for the positive assessment of our work, useful comments, and proposed improvements. Please find the answers to the specific questions below.

Answer to reviewer # 2

It would be helpful to also provide the derivation for Equation 3.

The derivation is straightforward and will be included in the Appendix of the revised version.

1. What are some of the current limitations of the paper? Is the learning system applicable directly to other types of optical interferometer alignment tasks? It would be nice to have some additional discussions in general.

The proposed learning scheme is applicable to any MZI as well as other types of optical interferometers (such as the Michelson interferometer) in which interference patterns can be extracted and subsequently used as a visual input to the agent. We will add a discussion to this effect to the Summary.

2. What is k in Eq 4?

It is the length of the optical wavevector with the components (k_x,k_y,k_z), i.e. $k = \sqrt{k_x^2 + k_y^2 + k_z^2} = 2\pi/\lambda$, where λ is the wavelength. We will include this missing definition in the revised version.

3. Line 143: should be Eq 3 instead of Eq 2?

In the real experiment, we do not have access to the internal state of the interferometer, which is described by the relative position and angle of the two beams. Therefore we cannot use Eq. (3) to calculate the visibility. Instead, the visibility is calculated via Eq. (2) from the interference pattern, which is known to the agent.

Answer to reviewer #3

Although it is reported that the human expert operated through a keyboard interface, I do not know whether it is a common way to align MZI in the field of optics. If it is significantly different from a usual way to align MZI, the performance of the human expert cannot be properly evaluated using the keyboard interface. For this reason, I cannot judge whether the claim "the robotic agent does outperform the human." is correct or not.

In the paper, we demonstrated two human expert benchmarks: using a keyboard interface with the same set of actions as the agent and manually using mirror knobs, in line with normal experimental practice. The results are shown in Fig. 5(a,b). The agent's policy outperforms human experts in both settings.

Although I understand practical benefits of the proposed system, I do not understand the difficulty of learning a policy for the automatic alignment of MZI.

Beyond the nontrivial RL problem setting of interpreting time-dependent interference pattern images and extracting optimal actions, our task is complicated by a number of factors associated with a real robotic setup. This includes (1) low rate at which the training data can be acquired; (2) pixel noise of the camera, leading to errors in evaluating the state and visibility; (3) uncertainty of actions: the angle by which a mirror is turned may differ from that specified; (4) day-to-day variations in the laser alignment, beam shape, ambient illumination and other conditions of the experiment, making it impossible to perfectly simulate the experiment.

The authors should discuss the recent studies on RL methods and domain randomization. Although the authors used dueling double DQN, there are some more techniques to accelerate DQN.

We agree that additional refinements to DQN, such as those implemented in the Rainbow algorithm cited by the Reviewer, could further improve the performance of Interferobot. At the same time, we observe that the double dueling DQN realized in our work is sufficient to achieve our goal of reaching high performance levels.

The papers on domain randomization that are mentioned by the reviewer are already cited in our manuscript.

Answer to reviewer # 6

I miss a discussion of what could be improved, if any.

We will add such a discussion in the revised version of the paper. In brief, future work will include (1) extending the method to other types of interferometers and interferometers with added optical elements such as lenses and (2) improving training algorighms with the vision of enhancing the sample efficiency to the extent that the policy can be trained without resorting to pretraining in a simulated environment.

A comparison to a black-box optimization approach (e..g CEM / evolutionary strategies) [...] would be very informative to the reader

We agree with this comment and will do our best to add this comparison to the paper before the revision deadline.

Answer to reviewer # 7

I could imagine one attempting other approaches to the problem: particularly a black-box optimization approach with compressed gradient sensing.

Please see our answer to Reviewer # 6 above.