(4) **I would like to read the following numbers in the paper: the size of the tweezers and the minimum step size (the expression lambda*f/mL is given in the paper, but one cannot compute it as L is not given). I feel they are important for the reader to visualize quantitatively how small are the changes between 2 holograms. Currently, only a qualitative comment is given ("overlap between consecutive tweezers").**

This is a good point. We specified dFFT = 0.45um and Gaussian beam waist of ~0.88um in the paragraph that mentions the formula of the Fourier unit.

(5) **Is there laser-cooling during the re-arrangement? If yes, this should be mentioned explicitly. If not, could this help to make larger steps and thus reduce the re-arrangement time?**

There is no laser cooling during rearrangement. Only in the start and final pattern do we perform laser cooling. It may be hard to perform cooling during transport in our non-magic traps. Although the linear interpolation method minimizes intensity flicker due to interference, the traps are still deformed during transport and consequently, the trap depth and imaging frequencies will vary. In non-magic traps, this is exaggerated with varying differential Stark shifts that make it hard to apply Sisyphus cooling.

Taking larger steps to reduce rearrangement time is very interesting. In the current implementation, we chose to have a Fourier unit that is significantly smaller than a Gaussian waist to reduce the deformation. Having constructive interference between two neighboring spots, it seems from calculations of the E-field that also for bigger steps (at least up to a Gaussian waist), a clear continuous trap is formed. An experimental study of how far one can go in step size would most likely include changing either the tweezer beam size or the Fourier unit to get enough data points, which is beyond the scope of the current work.

We have added a sentence in paragraph 5 of section “4. Scaling up to large atoms arrays”:

> “Additionally, one can consider moving atoms over distances greater than a single Fourier unit, which could greatly reduce the total number of holograms. However, we leave this for future work.”

(6) The re-arrangement time is dominated by the SLM refresh rate, but 30% is from the memory transfer. Could the authors comment briefly if this could be improved by implementing Direct Memory Transfer between the GPU and the SLM (bypassing the CPU)? What is the limit set by the PCIe bus bandwidth?

Indeed, we think that Direct Memory Transfer could improve the memory transfer benchmark, which is a significant portion of the total calculation time now. To obtain a limit set by the bus bandwidth we consider the specifications of our computer:

- Motherboard: Gigabyte Z790 AORUS MASTER | Supports PCIe 5.0x16
- CPU: Intel Core i9 13900K
- GPU: Gigabyte GeForce RTX 4090 24GB GDDR6X - PCIe 4.0

The GPU runs PCIe4.0x16. That means 16GT/s with 16 lanes, giving an absolute maximum of around 32GB/s. From this, we expect a bus limited duration of 33us to transfer a 1024x1024 array. One bottleneck will come from the PCIe bus used by the SLM, which is PCIe4x8. That is a limit of 66us to transfer to the SLM. This gives an order of magnitude minimum transfer time given by the current hardware, although we cannot comment on whether the SLM firmware would not introduce additional bottlenecks. While some time could thus potentially be gained with direct memory transfer, it should be noted that the total cycle time is still dominated by the refresh rate of the SLM. Therefore, we did not explicitly mention it in the paper.