Thank you to all reviewers for their thorough feedback. We're pleased that reviewers appear to endorse the overarching idea of the MAGICAL benchmark ("working on better IL benchmarks is a great idea", "sorely needed", "a lot of aspects have been taken into account"). We're also happy that reviewers found the writing clear and our claims correct. Reviewers raised several valid technical concerns and presentation issues, which we will address below.

Reviewer 1: It's good to hear R1 believes that "working on better IL benchmarks is a great idea". The Rubik's Cube ADR paper that R1 mentioned (Akkaya et al.) is relevant to MAGICAL, and we will include it in the camera-ready. Further, we hope that the following points will address the three outstanding issues R1 raised with the paper:

(1) "It is not clear to me whether [MAGICAL is] evaluating imitation learning (IL) or robust imitation learning (robust IL)." — Our aim was to measure the ability of IL algorithms to generalise far from the observed training data. The term "robust imitation learning" is a good way to avoid terminological confusion, and we will adopt it in the final revision.

(2) "... it is useful to argue that the new benchmark induces a reasonable ordering on algorithms ..." — This is a good point. We ran two of R1's suggested algorithms (Wasserstein GAIL-GP, and apprenticeship learning on autoencoder features) on a subset of three tasks (with fewer seeds and time steps), and show the results in Table AR1. We will include full results in the camera-ready. We did not find that either method generalised better than GAIL. This is perhaps to be expected: most existing IL algorithms do not specifically aim to improve generalisation (or robustness), and so it is not clear a priori what would constitute "reasonable ordering" of methods for this generalisation benchmark. We originally focused on multi-task variants, different views, and augmentation ablations in our experiments because we believed those axes would directly affect the ability of the network to generalise far from the training distribution.

(3) "Performance ... on the proposed benchmarks has a huge level of noise (Table 1)." — The standard deviations shown in Table 1 are pooled across all *tasks*, in addition to being pooled across all seeds. Most reported variance is thus due to the different task difficulty levels. Tables 5–8 in the supplement show that variation across seeds for each algorithm is much lower for most individual tasks. We will be careful to emphasise this in the final revision.

Reviewer 2: Thank you to R2 for their positive review! We are happy R2 agrees that MAGICAL "provides a benchmark which is sorely needed for IL". Regarding related work, Yu et al.'s Meta-World benchmark is indeed relevant and we will add it to Section 5. We will also trim the overlap between related work and the introduction, and devote less space to multi-task experiments, as requested; this will help free up space for the baselines requested by R1.

Reviewer 3: We're glad that R3 likes the idea of the paper and believes that the design process and methods are sound. As we understand it, R3 has two outstanding concerns:

(1) First, R3 raised concerns over our remark on lines 119–123 regarding tasks being "solvable by existing IL algorithms". We would like to emphasise that we only designed the *demonstration* variant of each task to be easy to solve. This allows researchers to focus on the fundamental challenge of MAGICAL, which is generalisation to the *test* variants. Tables 5–8 in the supplement show that existing algorithms fail to solve many of our test variants, even when they can solve the demonstration variant (e.g. the Layout variant of MatchRegions, ClusterColour, etc.). This presents a clear challenge to the community, which we believe is a prerequisite step to more ambitious tasks.

(2) Second, R3 suggested that "a better approach could be to make the existing frameworks easier to use ... rather than making a new one." We agree that ease-of-use and standards are important: for this reason, our implementation exposes each MAGICAL task and variant as a separate Gym environment, which should be easy to integrate with existing code. However, beyond the API, we do not think that small adjustments to existing benchmarks would suffice to evaluate generalisation. For instance, Section 5 notes that while it's possible to create "test" variants of Gym MuJoCo environments (e.g. the disabled ant), the underlying goal is still very simple, which limits how many distinct test variants can be created (there are only so many ways to run in one direction). In contrast, the object manipulation tasks in the MAGICAL benchmark have more complex goals that allow us to test many different axes of variation. We believe this justifies the introduction of a new benchmark, particularly given how few IL-specific benchmarks there are today.

Table AR1: Preliminary results for new methods on a subset of tasks (three seeds, 500k time steps per run).

	Demo	Jitter	Layout	Colour	Shape	CountPlus	Dynamics	All
				MoveToCorner				
GAIL	0.43±0.37	0.36±0.31	-	0.35±0.32	0.42±0.38	-	0.44±0.36	0.30±0.20
WGAIL-GP	0.04±0.03	0.05±0.05	-	0.01±0.01	0.08±0.10	-	0.05±0.05	0.01±0.01
Apprenticeship learning	0.00±0.00	0.00±0.00	-	0.00±0.00	0.00±0.00	-	0.00±0.00	0.00±0.00
				MoveToRegion				
GAIL	1.00±0.00	1.00±0.00	0.54±0.14	0.45±0.20	-	-	1.00±0.00	0.26±0.08
WGAIL-GP	0.97±0.01	0.97±0.01	0.47±0.02	0.25±0.02	-	-	0.97±0.02	0.14±0.01
Apprenticeship learning	0.33±0.47	0.33±0.46	0.10±0.14	0.09±0.12	-	-	0.33±0.47	0.05±0.07
				FixColour				
GAIL	1.00±0.00	0.67±0.11	0.21±0.03	0.32±0.03	1.00±0.00	0.16±0.04	0.97±0.04	0.14±0.04
WGAIL-GP	0.65±0.42	0.31±0.21	0.06±0.02	0.10±0.04	0.66±0.40	0.02±0.03	0.66±0.37	0.04±0.03
Apprentice. learn.	0.01±0.01	0.00±0.00	0.01±0.01	0.05±0.02	0.03±0.02	0.00±0.00	0.01±0.02	0.01±0.00