We thank the reviewers for their valuable feedback and encouraging reception of our work. We're glad they found our 3D object-centric model of videos to be highly significant, of interest to the NeurIPS community, and potentially impactful. We now address some points raised in the reviews; we will of course incorporate the other suggestions.

R1: videos are short; illustrate 3D with different cameras. We have now trained a model on 12-frame (ROOMS) videos, showing our method can scale to significantly longer sequences. Fig. A shows generated images, objects, and depths, which still show coherent scenes. Fig. B shows reconstructions; the rows are original, reconstructed, reconstructed with higher camera angle (to better show the 3D structure), and objects with 3D bounding boxes (from that angle). The additional images to the right use even more extreme viewing angles, revealing the 3D layout clearly.

R1: low resolution & fidelity. Training video models is expensive, particularly in 3D. We consider the resolutions used (96x72 & 80x80) a good trade-off between computation and quality; they are also comparable with similar works [10,11,21]. Our results show O3V-voxel produces higher-quality videos than the state-of-the-art (SCALOR) on the complex (TRAFFIC) dataset (Tab. 2b, Fig. 5 & S6). Moreover, our method is the first that can address our tasks and setting; it is natural that scope remains for improving visual fidelity and resolution in future work.

R1,2,4: complexity of datasets. Our claim of handling more visually-complex videos than prior work was meant with reference to the state-of-the-art in generative object-centric video modelling (SCALOR) [21]; we'll clarify this. [21] is demonstrated only on objects of near-uniform color, without shading nor perspective/3D effects. O3V successfully models videos containing all these effects, while SCALOR fails to do so (Fig. S6 & Tab. 1). While we do only demonstrate O3V on synthetic videos (**R2**), our (TRAFFIC) dataset contains significantly more complex videos than state-of-the-art (**R4**)—and so we already go a significant way towards bridging the gap to natural videos.

R4: limited number of object slots G. Many similar models have the same limitation [21,10,29], but we agree it would be interesting to lift it. Note G is rather large in our case (tens of objects).

R1: correctness/clarity of L86-87. Our intended meaning is simply that the grid is 3D, with its cells placed in 3D world-space, rather than being a 2D grid in the space of the image. Such a 2D grid is used in the generative models [6,21], inspired by YOLO [34], which produces detections based on a grid of cells. We'll clarify in the camera-ready.

R1: overly strong claims; inductive biases are 3D supervision. We note that numerous related models (e.g. SQAIR, SCALOR, IODINE), regarded by the community as unsupervised, have similar inductive biases (priors on object size, speed, uniformity of color), albeit in 2D. We therefore respectfully disagree with this characterisation. Of course, we do agree that inductive biases make learning easier (indeed, possible!). We will qualify the claim of compositionality at L36 to note this refers to the generative model itself, not the encoder used to train it.

R4: explain differing generation quality. We'll expand the discussion in Sec. 5.2. O3V-mesh has poor FID on (TRAFFIC) as it is prone to local optima where cars are not tracked correctly—see the images in the supplementary. GENESIS has good FID on (ROOMS) as color segmentation (which it readily exploits) is a strong cue here.

R1,2: use of FID. We agree KID is a more-principled generation metric than FID; we used FID for consistency with the works we compare to. Following **R1**'s suggestion, we re-ran our generation evaluation using KID (see Tab. A). We see that all statements made in the paper regarding relative quality of methods remain true—in fact, FID & KID are highly correlated on our datasets. Note (**R2**) that in calculating FID, the ground-truth feature distribution is defined by synthetic images, so the use of a 'natural image network' is reasonable (and, indeed, common [10,28]).

R2,3: several components of O3V (e.g. mesh renderer) are known techniques. This is true for very many works, and there is clearly value in showing that a novel model using some existing techniques can achieve state-of-the-art results. Importantly, we also examine different variants (i.e. mesh vs. voxel representations), and discuss their strengths.

R1: authors do not show scene generation results from prior works. These are in supplementary figures S1–S6.

R3: motivation for discrete grid of objects. We'll expand the discussion at L90. A grid of objects ensures gradients with respect to object locations are non-zero, even when the current predictions are poor, as at least one candidate object should be near enough each true object. Note that the actual object locations are continuous, as they are offset by a vector Δ_g from the cell center.

R1: compare to GQN and SRN. These perform novel-view synthesis, but cannot sample new scenes *a priori*. They also cannot perform segmentation/tracking, as they lack a representation of separate objects and assume a static scene.

	(ROOMS)	(TRAFFIC)
MONet	0.151	0.305
GENESIS	**0.083**	0.257
SCALOR	0.148	0.272
O3V-voxel	0.108	**0.157**
O3V-mesh	0.106	0.345

Table A: KID scores

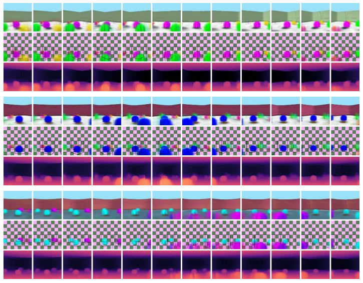

Fig. A: 12-frame generations

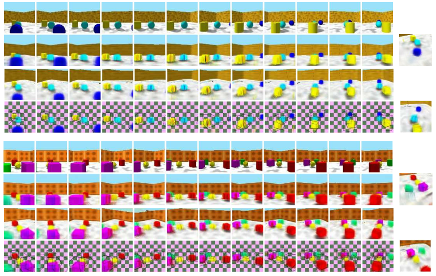

Fig. B: 12-frame reconstructions