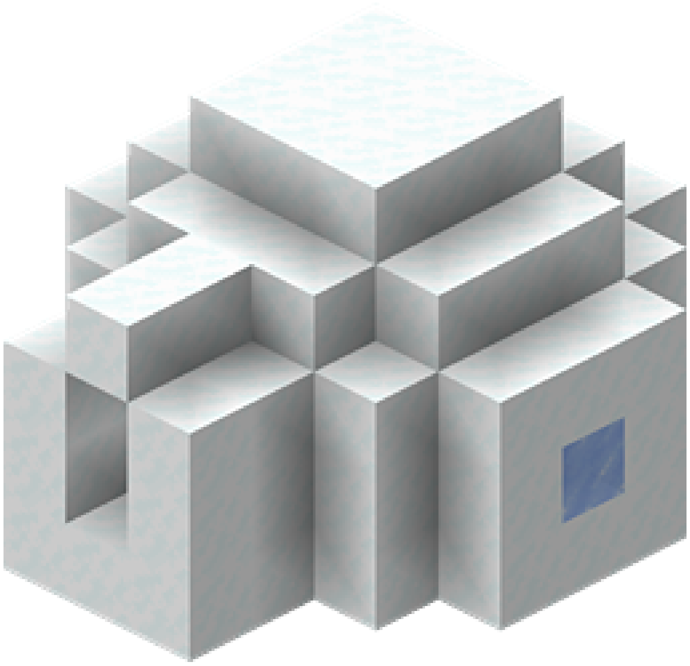


NeurIPS 2022 - IGLU Challenge - Chuang submission

Overview

We are seeking the research award. The majority of our time was spent on engineering a novel 3D diffusion architecture as the high level goal generator rather than the T5 module. Although our internal evaluation was very promising, we ran out of time for making this submission work well for the external public evaluation, and achieved only a .09 F1 score with our diffusion model. The specifics on why it did not work well are detailed below. As a result, out final and highest scoring submission is *just the baseline with no modification.* However, we believe that the work done on the diffusion model is still of research interest and is a valuable contribution to the Language+RL community. Therefore, we detail our diffusion work for research award consideration.

3D Discrete Diffusion with Context Prompting

We build off of the MHB baseline and attempt to replace the T5 model with a better goal generation module. We first make the novel observation that attempting to generate the goal voxel grid from text can be reformulated as a text to video problem, where the time dimension in the video is reinterpreted as the third dimension in the voxel grid. Inspired by the recent success of the diffusion model Imagen for highly detailed and novel image generation, we adopt an open source implementation of video diffusion models for the IGLU 2022 challenge and make the following key contributions:

- Context prompting through classifier guidance and 3D reconstruction
- Continuous to discrete generation through thresholding

We instantiate the diffusion model as a 3D Unet with classifier free language guidance provided by a frozen pretrained T5 encoder. We use the standard denoising objective with the provided singleturn and multiturn dataset for training. We define the `delta grid := target_grid - starting_grid`. Furthermore, we change the value of all unoccupied voxels to -1, and occupied voxels to 1. Our model attempts to predict this delta grid from the language instruction. We defer to this well written article for background details on the vanilla Imagen model.

Context prompting

Critically, we prefix each of our language instructions with the initial starting grid. For example, if the language instruction is `put a block to the right of the current center block`, and there is one orange block at (x0, y0, z0), we will feed `There are blocks on z0,y0,x0,orange. Implement: put a block to the right of the current center block` to our language encoder. Since many of the instructions provided in the dataset are not comprehensible without knowledge of the context and current blocks in the environment, this language prompting enables contextual generation and encodes the necessary data for our diffusion model to effectively denoise to the target delta grid.

In order to retrieve the starting grid during evaluation, we hard code the agent to walk around the environment once, and perform a multiview 3D reconstruction from the agent's image observations. This hard coded behavior was explicitly given permission by Alexey Skrynnik.

Thresholding

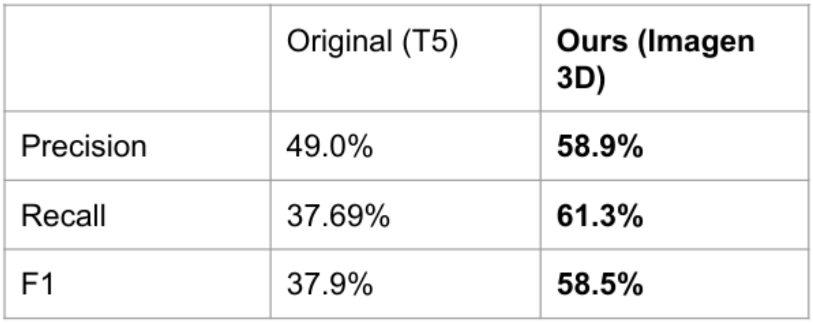
An issue that is immediately apparent in applying off the shelf diffusion models is that they operate in a continuous space. Whilst diffusion models in discrete space have been explored, we were not able to find any research that considers both discrete and 3D space. To the best of our knowledge, in this work we present the first model that can do both, through a simple but effective technique. Our diffusion model will generate a 3D voxel grid from text, where the values in each cell of this grid are interpreted as the log probability of a block existing at that location. By thresholding the output of our model at t=0, we simply clip all values less than t to 0, and all values greater than t to 1. This approach can be justified by a theoretical analysis, as the denoising MSE objective used here is nearly equivalent to a shifted pairwise binary cross entropy loss between the predicted logits and target grid for each voxel. Through the denoising process, our diffusion model slowly converges to the true occupancy distribution for each individual voxel, whilst still robustly modeling the complex join dependencies between voxels through usage of 3D convolutional filters in our Unet.

Summary

Using contextual 3D diffusion models for IGLU yields a number of benefits. Firstly and most importantly, the performance is significantly improved over the T5 model. Secondly, it enables a natural method for the precision and recall tradeoff through the diffusion forward process. The forward process (adding noise to the starting grid) is equivalent to sampling from a more entropic distribution, thereby increasing recall and decreasing precision. Since this diffusion process is done in continuous time, it provides a natural lever for trading off recall and precision, enabling a black box optimization of the F1 score through finding an optimal timestep to diffuse forward to. This allows a highly configurable and tunable setup, ideal for the evaluation benchmarks of the IGLU challenge. Finally, it holds better inductive bias for this problem as it generates all blocks simultaneously, which allows the capturing of long range dependencies between blocks and the enforcing of global consistency. This is something the original T5 decoder is not capable of.

Performance Evaluation

We present our scores for the full set of training tasks compared against the T5 baseline, where we improve on the baseline's F1 score by **54.3%**. We calculate these scores by using the provided code in the starter kit, which finds the maximal intersection between the ground truth delta grid and predicted delta grid.



	Original (T5)	**Ours (Imagen 3D)**
Precision	49.0%	**58.9%**
Recall	37.69%	**61.3%**
F1	37.9%	**58.5%**

Figure 1: Our scores against the original

Below we additionally present our scores for our diffusion model during training on a sample of 100 of the 548 training tasks provided. Adding context prompted diffusion is very effective, reaching a F1 score of .81 at 230K gradient steps. We truncate this graph as we did not collect scores before 120K steps.

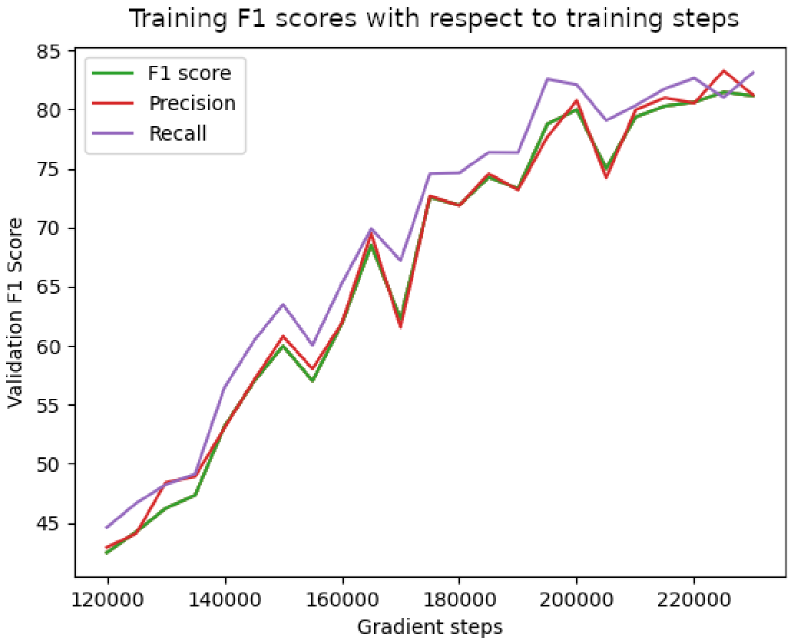


Figure 2: Training F1 scores with respect to training steps

Why didn't diffusion work for the full competition evaluation?

Most critically, we were not able to construct a robust 3D reconstruction algorithm in time for the final submission. This meant we were not able to prefix an accurate context prompt for our instructions. Methods we attempted for the 3D reconstruction included a multi-view transformer (Legoformer), a fast neural representation field implementation using voxels (Plenoxels), and a simple CNN-MLP. In the end, we did not have enough time to get any of these methods to work well. The best performing was the CNN-MLP with a validation F1 score of around .5, which was too low for our model to work well during evaluation. A promising future direction is simply using traditional computer vision with point-correspondences, as the extrinsic and intrinsic camera matrices for each image used during reconstruction is known apriori.

Code details

You can view our submitted agent here and the submitted diffusion model code here.