Dear reviewers,

we want to thank you for taking the time to read through the latest draft and provide thorough feedback! We have adapted the survey to address the concerns that were raised and will detail them here.

We have expanded our motivation for the ASP focus from a few sentences to three paragraphs (lines 187 to 217). We give three main arguments as to why ASP is distinctively worth surveying: First, the stable model semantics allows frameworks to consider multiple worlds at once, but also increases the computational complexity and difficulty of assigning gradients. Second, the non-monotonicity and default rules are well-suited for modelling natural language logic problems. This is particularly important for all the LLM-based papers, which make up a large part of the survey and have often cited these capabilities as the reason for choosing ASP. Third, there is a mature ecosystem of efficient software for ASP, both for solving and learning. The last two points in particular are completely new arguments in this latest version. And as you mentioned, at some point there needs to be a cutoff for which papers to consider, even if it is somewhat arbitrary.

Regarding the performance tables, we agree that the large number of empty spaces limits their usefulness. We have therefore split up the first table into four separate tables (no. 2-5). This removes a lot of whitespace and allows the reader to focus on the specific frameworks we are currently discussing in each section. The second table has been reduced to only the frameworks with shared datasets. We have integrated the remaining accuracies into the text. We also agree in principle that extending the comparisons would be a valuable contribution, but this would require a large amount of effort. This survey focusses more on describing the landscape, rather contributing new datapoints. Since the meta reviewer has said that it is not necessary to run any experiments ourselves, we will stick to that.

For the running times, we have followed your suggestion to display the data in graphs, which we agree is much more intuitive. Only two sets of graphs are needed to depict almost the entire table: Figures 8 and 9. The former is a line graph, as the tasks include multiple versions that increase in complexity. The latter is a grouped bar chart showing the scale of the differences in runtime. The remaining two datapoints have been integrated into the text.

We have also tried to separate the perception task description from the analysis by moving the dataset descriptions to their own section. Unfortunately, this did not work very well. The majority of sentences in the Perception tasks section are analytical, discussing dataset difficulty and latent label availability. We have attached an image of a snippet of this section on the next page, highlighting the analytical sentences in green.

attributes into one predicate. While the perception task with ShapeWorld is more difficult, the actual neurosymbolic task is easier than for MNIST tasks.

In the CLEVR dataset, the shapes are three-dimensional and can partially occlude each other (Johnson et al. 2017). Again, SLASH trains the neural component directly on the latent attributes of the objects. ASP-VQA and AQuA use a pre-trained YOLO network and do not perform any learning at all. NeSyGPT uses a VLM instead, which is pre-trained on a large corpus of general images. In addition, the authors finetune it with a small number of latent labels. Similarly, the CLEGR^V dataset generates images of graphs deterministically from their specifications. This results in limited variety and a straightforward perception task, which NSGRAPH solves using pre-trained graph recognition models.

In the category of synthetic images, only MNIST is truly used for neurosymbolic training of the perception component. Most MNIST tasks include just the downstream labels and therefore only provide noisy signals for the latent classification task. ShapeWorld, CLEVR and CLEGR^V, while embodying a more complex perception task, provide direct latent labels in all cases.

Real-world images. CIFAR-10 is a dataset of real-world images compressed to 32x32 pixels that depict one of ten object categories (Krizhevsky and Hinton 2009). Embed2Sym uses it for the CIFAR-10 Addition task, where each image category is arbitrarily assigned a number and the goal is to find the sum of two images. Just like in MNIST Addition, no latent labels are given, but the perception task is more difficult.

The Visual Question Answering and Reasoning (VQAR) dataset contains diverse real-world images with a much higher resolution than CIFAR-10 (Huang et al. 2021). But the SLASH framework uses a pre-trained network to find the bounding boxes of objects, sidestepping the difficult task of training the neural component.

The Playing cards dataset consists of photos of real playing cards with a resolution of 523x831 pixels (Cunnington et al. 2023a). It is used to learn the rules for determining the winner of card games. None of the three papers that use the dataset actually train the neural component from downstream labels. FFNSL pre-trains the neural component on the latent playing card labels directly, while Embed2Rule and NeSyGPT use a VLM with latent finetuning.

The same is the case for the PlantVillage (Hughes and Salathe 2016) and Indoor scenes (Quattoni and Torralba 2009) datasets. While they contain real-world images of diseased crops and varied indoor rooms, FFNSL uses a pre-trained neural network. For the PlantDoc dataset, which contains images of diseased plants (Singh et al. 2020), NeSyGPT uses a VLM with latent finetuning.

The dataset catalogue ended up being a very dry and repetitive description of the datasets. When rewriting the analysis, we had to repeat these descriptions anyway to provide the necessary context. We believe a separation would worsen the flow of reading and repeat information unnecessarily. Therefore, we have decided to keep the section as it was. We hope our rationale makes sense!

Thank you as well for the smaller comments. We have included GNNs in the background section (lines 361-365) and corrected the grammatical errors.