Thank you for the detailed and constructive comments. Following the reviews, we conducted the following experiments:

1. We ran our method on RAVEN-FAIR [1], see the Fig. I. Note that some attributes are allowed to change when no rules are applied on them. As noted by [1,B], the original RAVEN [A] is biased and CoPINet [13] exploits this. Since CoPINet does not perform as well on unbiased data, we evaluated using MRNet (SOTA model with 80.6% accuracy) instead. The generation accuracy was 61.2%, this is to be compared to 66.8% on the target image reconstructed by VAE, and only 9.0% on random generated image.
2. We conducted a user study, following the same scheme as the machine evaluation. We extensively trained three participants on the task of PGM questions. After training, each got 30 random questions with the correct target image (reconstructed) and 30 with the generated target instead. Human performance on the correct image was 72.2%, and on the generated image was 63.3%. This small gap reassures that the generation is accurate.
3. A second user study was appropriate for untrained humans. This study is similar to the qualitative image analysis shown in the paper (Fig. 4). In PGM, an image is correct if and only if it contains the right instance of the object attribute which the rule is applied on (this information is in the metadata). By comparing the generated object attribute to the correct image's object attribute – it can be easily determined if the generation is correct. The study has 22 participants, 140 random image comparing instances for the generated answers, and 140 for a random choice image (reconstructed by VAE) as a baseline. 70.1% of the generations were found to be correct and only 6.4% of the random choice images (baseline). Considering that the SOTA model in the much easier task of recognition achieves 75.2% (MRNet), it seems that the generation accuracy does not fall much behind.
4. To demonstrate generalization, we trained on the "interpolation" regime of PGM in which the rules of train and test differ. We evaluated using MRNet that was trained on the "neutral" regime (we measure the generalization of our method, not of MRNet) and received 54.3% generation accuracy, this is to be compared to 70.8% accuracy on real targets. This ability to generate abstract images based on patterns that are deliberately different than those of the training set is quite remarkable.

Reviewer 1: We made the evaluation with multiple models simultaneously to not base the conclusion on a single model that might be biased. Following the review, we added experiments with RAVEN-FAIR and two user studies (see points 1-3 above). As noted in point 1 above CoPINet's success is partly due to the ability to answer RAVEN questions without looking at the context (the query).

Reviewer 2: W1: According to Sec.A.2 in the appendix of [10], human performance on PGM is very low. However, very experienced participants scored well (80%). In the RAVEN dataset humans tend to score roughly 84% correctly on average [A]. **W2:** We disagree that we just combine VAE and GAN in this work. VAE-GAN is used to an unconditional generator, but this is only a starting point. In L147-151, we add a novel recognition path that selects a vector in the VAE latent space. A novel relation-wise perceptual loss is defined (L173-196). These, in addition to the novel CEN model in L100-103 that facilitates this, are the main contributions in this work. **W3:** As mentioned in L59-60, we use the PGM [10] dataset. Following **R1**'s request we also use RAVEN-FAIR.

Reviewer 3: Untrained humans score low on PGM (see above); other PGM-based tasks also require supervised training. This is similar to, e.g., computer chess. **Generation:** As we discuss, a multiple-choice protocol is easily exploitable. In contrast, (1) generation is much more likely to require reasoning and (2) the important problem of abstract generation was not studied in the past, as far as we can ascertain. **DS-KLD:** DS-KLD is designed to create variance in a subset of the vector's indices. This subset is dynamic and depends on the nature of each vector. It could be beneficial for any problem for which there are multiple modes to the latent representation and these modes are anisotropic. **User study:** The 'blurry patterns' mentioned may not be a problem for a study if all choices are reconstructed by VAE. This way users cannot pick on reconstruction artifacts, if these exist (this is why we apply such reconstruction in the user studies in points 2+3 above). Note that humans without experience on PGM tests perform very poorly (see response to **R2**), therefore it is hard to create a comprehensive study without investing considerably in training users. To circumvent this limitation, we have also performed a user study that is suitable for untrained individuals. **Split:** We trained on the train set and evaluated on the test set of PGM (and the same for RAVEN-FAIR). **out-of-distribution:** Following the remark, we trained on the "interpolation" regime of PGM (see 4) and the results show very convincing generalization.

Final note: All reviewers have highlighted that the paper is well written, the idea is novel and interesting, agreed that the results were good and the ablation study was extensive. We hope the reviewers would be able to read our rebuttal with an open heart. Thank you.

Additional references:

[A] Zhang et al. RAVEN: A Dataset for Relational and Analogical Visual rEasoNing. CVPR 2019.

[B] Hu et al. Hierarchical Rule Induction Network for Abstract Visual Reasoning. arXiv 2002.06838, 2020.

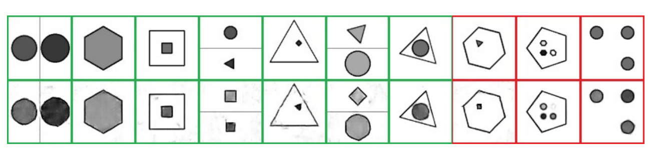

Figure I: **Top:** A correct RAVEN-FAIR answer. **Bottom:** The generated answer (correct, even if differs, in green, incorrect in red).