We thank the reviewers for their insightful and valuable feedback and for their *unanimous support* of the paper. We are encouraged that they found our formulation to be **"novel/new"** (**R1**,**R2**,**R3**) **"interesting"** (**R2**,**R3**,**R4**) and with a **"strong direction"** (**R4**); The experimental results to be **"effective"** (**R1**,**R2**,**R3**,**R4**), **"extensive"**,**"thorough"** (**R2**,**R4**), **"detailed"** (**R2**,**R3**) and **"strong"** (**R4**); And AO-CLEVr dataset **"beneficial"**,**"good"** (**R2**,**R3**).

R4: DAG is most suitable to disentangled attributes, some attribute manifest differently depending on object. Extending to dependent pairs is a very important next direction. Unfortunately, it is not clear that zero-shot can work well with strongly entangled pairs because every case could be special. Three insights worth mentioning: (1) Even the fully disentangled case is still very challenging. (2) The factored DAG can be used as a strong zero-shot prior for few-shot learning, thus benefiting future work on dependent pairs. (3) Using a "closed" settings may capture some of the dependency by eliminating "over-generalization" (e.g. by disallowing yellow-wine label).

R4: Clarify details of mapping the causal graph to Fig 1c (1) The role of g_A^{-1}, g_O^{-1}, do they add assumptions? context to causal DAG? g_A^{-1} and g_O^{-1} are used to estimate the latent ϕ_a and ϕ_o of an image instance. They reflect assumptions about the noise level in the data-generation process (Suppl L508-513), i.e. that the mapping from the core-features (ϕ_a and ϕ_o) to the image ($\mathbf{x}$) is not too noisy and the latent vector can be recovered from the image. (2) The new nodes $\hat{\phi}_a, \hat{\phi}_o$ satisfy the independence constraints by construction. Explain consistency. We respectfully point out that since $\hat{\phi}_a, \hat{\phi}_o$ are children of $\mathbf{x}$, they *do not* satisfy the independence constraints of Eq. 6. Minimizing L_{indep} encourages the property $p(\hat{\phi}_o|do(o)){\approx}p(\hat{\phi}_o|do(a,o))$ (L550). Only then the independence relations of Eq. (6) apply to $\hat{\phi}_a, \hat{\phi}_o$. It also minimizes the PIDA metric of (Suter 2019). (3) Any assumptions fail for MLPs? No.

R4: Where does the causal interpretation manifests? (1) Difference from standard embedding? As the reviewer points out, one difference is in the independence loss; another is the use of two separate embedding terms tied to the independence loss; both motivated by the causal graph. We deliberately proposed a model close to baselines (L186) to surgically demonstrate the strength of the proposed approach. (2) Why is λ_{invert} essential? Since there exist no ground truth values for neither ϕ_a nor h_a, minimizing $||\hat{\phi}_a - h_a||^2$ may reach trivial solutions (same for ϕ_o, h_o). λ_{invert} guides the optimization and avoids trivial solutions. It does not contradicts assumptions on the causal process.

R1: How are the means $h_a, h_o, g(h_a, h_o)$ updated? Instead of learning explicit values for the means, we learn MLPs that output the means (using gradient updates L138,143,178). For example, an MLP (h_A) maps the (one-hot) representation of "leather" to $h_{leather}$ and an MLP (g), maps $(h_{leather}, h_{sandal})$ to $g(h_{leather}, h_{sandal})$.
R1: Does interventional inference means matching prototypes? Partially yes: Inference that follows *the approximations we took* (Supp A., e.g. Gaussian and 0^{th} order Taylor) may be viewed as matching prototypes. In the general case, there may be better ways to estimate the likelihood of $p(\mathbf{x}|a,o)$ and the factors $p(\phi_a|a), p(\phi_o|o), p(x|g(\phi_a, \phi_o))$.
R1: Disentanglement is achieved by independence loss rather than intervention. The independence loss allows to learn a model that is robust to interventions. Minimizing L_{indep} encourages $p(\phi_o|do(o)){\approx}p(\phi_o|do(a,o))$ (L550).

R2: Independence loss encourages the performance on the unseen data but drops on the seen data. This is a known and important trade-off (Rothenhäusler 2018): The independence loss discourages certain types of correlations, hence models do not benefit of them when the test and train distributions are identical. However, the loss is constructed in such a way that these are exactly the correlations that fail to hold once the test distribution changes (zero-shot). Ignoring these correlations improves performance on unseen data. We will refer to (Rothenhäusler 2018) and discuss.
R2: Failure analysis. Following this request, we analyzed samples of unseen pairs of Zappos in the open-world setup. We compared *Causal* with LE*, which is the strongest no-prior baseline. LE* confuses unseen pairs for seen pairs at a rate of $3.7:1$, while *Causal* errors are more balanced $1.2:1$. One interesting failure case of *Causal*, is that it over-commits for predicting the pair "Leather-Slippers", which was unseen during training. In the final version we will provide more qualitative and quantitative details about Zappos and AO-CLEVr .

R3:MIT dataset: Consider top-k labels: Following this suggestion, we conducted a new experiment to evaluate both top-1 and top-2 accuracy. Raters were asked to select the best and 2nd-best attributes that describe an image, among attributes relevant for that object. The top-1 accuracy was 32%, consistent with previous experiment. The top-2 accuracy was 47%, only slightly higher than adding a random label on top of top-1 label (yielding 42%). To verify that raters were attentive, we also injected 30 "sanity" questions that had two "easy" attributes, yielding top-2=100%.

R3: MI instead of HSIC: HSIC advantage is that it is non-parametric, unlike MI, and does not requires training an additional network for variational approximation. Embed. size; We will report results w.r.t. embedding size. Efficacy of HSIC: See λ_{indep}=0 at Table S.2. Results w/o alternate training. Alternate training lowers the SEM (L713). Means are comparable (68.7 vs 67.7).

R3: Revise method for smoother reading. **R4** Some bits are confusing. We will restructure the paper based on your feedback: (1) Shorten the "overview" section (2) Discuss how independence loss allows $\hat{\phi}_a, \hat{\phi}_o$ to recover the properties of ϕ_a, ϕ_o, and its relation to PIDA. (3) Update the final version based on the rebuttal.

R1,**R2**,**R3**,**R4**: We will address all minor comments, and clarify the broader impact.