We thank the reviewers for their constructive assessment and for the many positive remarks regarding high significance and relevance (R1+R3), technical contribution (R1+R2), correctness (R2+R4), clarity (R1+R2+R4), and discussion of limitations and impact (R1+R2), in addition to a perceived consensus that the experiments support our claims. Thus, we find **R1** and **R2**'s lower scores surprising, given their otherwise supportive statements. We kindly invite each reviewer to revisit their recommendation in light of our rebuttal, as most concerns are easily addressed with minor clarifications.

[R2+R3] Adopted definitions of counterfactuals etc.: Though we attempted to distinguish our usage of 'interventions' vs. 'counterfactuals' in L53–58, we acknowledge that these definitions are not universal and may differ across causality paradigms. Following Pearl [2] and Peters et al. [1], 'interventions' operate at the population level, providing aggregate statistics about the effects of actions (i.e. noise sampled from the prior, $P(\boldsymbol{\epsilon})$). In contrast, a 'counterfactual' is a retrospective hypothetical query at the unit level (cf. potential outcome), where the structural assignments ('mechanisms') are changed but the exogenous noise is identical to that of the observed datum ($P(\boldsymbol{\epsilon}|\mathbf{x})$). Further, note that an unseen combination of antecedents (e.g. 'woman+mustache' in [14]) is not a counterfactual under this definition. As suggested by **R3**, we will clarify our usage of these terms to avoid ambiguity in the final version.

[R1+R2] Clarification of our contribution and comparison to SOTA: **R1** and **R2** wished to see a comparison of our approach against level-2 SOTA methods in deep-learning and causal-inference literatures, respectively. However, as outlined in the title, abstract, and throughout the text, the principal claim of this manuscript is a novel approach to tractable *counterfactual inference* using deep-learning components, hence our focus on level 3 in the experimental validation. Although fulfilling the top rung of Pearl's Ladder of Causation does indeed entail generative and interventional capabilities—as demonstrated in the main text and supplement—here we make no claims of superiority of our approach on those two levels. We believe a comparison to level-2 methods is therefore beyond the scope of the current article.

[R2+R3] Assumptions, misspecification, and unobserved confounding: We agree that Markovianity/no-unobserved-confounders is a strong assumption and will further highlight this explicitly earlier in the paper, rather than only at the end. We shall also put more emphasis on the interpretation of our results under the light of misspecification of the simpler baseline models (**R3**). The current formulation indeed assumes causal sufficiency: every latent variable affects a single observed variable, and all relevant dependencies are captured in the DAG. Yet, to the best of our knowledge, this is the first successful attempt to build deep structural causal models (Markovian or not) that can produce high-dimensional counterfactuals. We appreciate **R2**'s suggestion regarding semi-Markovian models, and recall that our next steps involve the handling of missing variables, which is related to unobserved confounding.

[R3] Novelty and related work: The remark that CausalGAN [14] 'can create counterfactual images like "what if this person had a mustache?" ' is inaccurate, as such a query would relate to adding a mustache to an existing (*factual*) image, not to generating novel samples. That paper focuses entirely on interventions, regarding which it unquestionably makes a strong contribution—e.g. enabling sampling from the interventional distribution $p(\mathrm{image}|\mathrm{do}(\mathrm{mustache} := 1))$. However, this is not a counterfactual in the sense of a unit-specific retrospective intervention. The paper contains a single passing mention of counterfactuals (p. 8) with no corresponding experiments and assuming the exogenous noise is known—effectively avoiding to address the challenge of abduction like the other related works cited.

[R1] Soft interventions: Figure 3 already illustrates soft interventions on thickness of the form $\widetilde{f}_T(\cdot) = f_T(\cdot) + \tau$, with $\tau = +1$ and -0.5. We note that the key challenge in computing counterfactuals within Pearl's three-step procedure is *abduction*. Both the *action* (intervention) and *prediction* (final forward pass) steps are trivially computed once abduction has been performed, by simply overwriting the assignments with arbitrary functions.

[R4] Evaluation of counterfactuals: Following **R4**'s suggestion, we generated synthetic counterfactual images (with $\mathrm{do}(t+2)$) and computed the pixel-level mean absolute errors: the full model achieved 17.6, while the conditional and independent models reached 41.6 and 31.8, resp. For real datasets, counterfactual evaluation is possible only in very constrained settings, as generally true counterfactuals can never be observed. For example, assuming our brain graph is correct, we may use a second MRI scan of a brain a few years later as an approximate counterfactual on age.

Other clarifications: **[R1] Implicit-likelihood mechanisms:** This class of mechanisms was included for completeness; we only speculate about their feasibility and make no claims on performance. This will be emphasised in the text; evaluation is left for follow-up work. **[R4] Evaluation metrics:** The log-likelihood (and lower bounds) is the training objective being optimised (see L112–113 and L135–136); it measures the level-1 fitness of a probabilistic model to the dataset, and is comparable across different models and parametrisations. **[R4] Morpho-MNIST SCM:** We will further clarify that the model in Eq. (7) was designed by us and used to generate the synthetic dataset. **[R1] Brain SCM:** Although inspired by medical evidence, our brain SCM is admittedly oversimplified and meant only as a proof of concept. We will emphasise in the text that any conclusions drawn from a model built in this framework are strictly contingent on the correctness of the assumed SCM. **[R1+R2] Suggested references:** We appreciate the pointers to Tran & Blei (2018) and Louizos et al. (2017), which we agree will strengthen our discussion. Both introduce deep level-2 methods (implicit and amortised, resp.) that explicitly deal with unobserved confounding.